Use of SNP markers to conserve genome-wide genetic diversity in livestock

Krista Anika Engelsma

Thesis committee

Promotor
Prof. dr. ir. J.A.M. van Arendonk
Professor of Animal Breeding and Genetics
Wageningen University

Co-promotor
Dr. J.J. Windig
Senior researcher, Animal Breeding and Genomics Centre
Wageningen UR Livestock Research

Other members
Dr. M. Tixier Boichard, INRA, Jouy-en-Josas, France
Prof. dr. F.A. van Eeuwijk, Wageningen University
Dr. ir. T.J.L. van Hintum, Centre for Genetic Resources, the Netherlands, Wageningen
Prof. dr. ir. T.H.E. Meuwissen, Norwegian University of Life Sciences, Ås, Norway

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Use of SNP markers to conserve genome-wide genetic diversity in livestock

Krista Anika Engelsma

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday December 7, 2012
at 4 p.m. in the Aula.

Engelsma, K.A.
Use of SNP markers to conserve genome-wide genetic diversity in livestock.

PhD thesis, Wageningen University, the Netherlands (2012)
With references, with summaries in English and Dutch

ISBN 978-94-6173-386-3

Abstract

Engelsma, K.A. (2012). Use of SNP markers to conserve genome-wide genetic diversity in livestock. PhD thesis, Wageningen University, the Netherlands

Conservation of genetic diversity in livestock breeds is important since it is, both within and between breeds, under threat. The availability of large numbers of SNP markers has resulted in new opportunities to estimate genetic diversity in more detail, and to improve prioritization of animals for conservation of genetic diversity. The aim of this thesis was to further explore the potential of SNP markers for estimation and conservation of genetic diversity within livestock breeds. This was evaluated analyzing Holstein cattle populations, genotyped with a commonly used 50k SNP chip. Genetic diversity was estimated with SNP markers and compared to genetic diversity estimated with pedigree information. Both methods could detect differences in overall genetic diversity, even between two closely related populations. With SNP markers, differences in genetic diversity at the chromosomal level could be identified as well. Subsequently, SNP markers and pedigree information were used to prioritize animals for conservation in a gene bank using optimal contributions. SNP based prioritization was slightly more effective than pedigree based information, both over the whole genome and at specific regions of the genome. We extended the optimal contribution method to simultaneously conserve a single allele at a specific frequency and maximize the overall genetic diversity conserved in a gene bank. The loss of overall genetic diversity was larger when the target frequency for animals conserved in the gene bank differed more from the original frequency in the population. It can be concluded that dense SNP data form a powerful tool for estimation and conservation of genetic diversity in livestock breeds. Although pedigree information gives a good representation of the overall genetic diversity, SNP markers can provide more detailed information about the genetic diversity over the genome. Especially for small populations, SNP markers can play an important role in conservation of unique alleles, while simultaneously minimizing the loss of genetic diversity at the rest of the genome.

Contents

1

General introduction

1.1 Genetic diversity in livestock

Genetic diversity in many livestock breeds across the world is threatened. Decades of selection aimed mainly at production increase has led to the irreversible loss of genetic diversity (FAO, 2009). This loss can be disappearance of breeds. In countries all over the world local breeds have been replaced by high production breeds, such as the Holstein Friesian cattle breed. In Box 1.1, the breed risk status of the major livestock species in the world is given.

Box 1.1 Breed risk status in the world.

Of the in total 7616 livestock breeds in the world that have been reported in the FAO's Global Databank, about twenty percent of the breeds are classified at risk, and almost one breed per month was lost during the last six years (FAO, 2007a). Especially local breeds are threatened with extinction. In figure 1, breed risk status for the major livestock species in the world is shown.

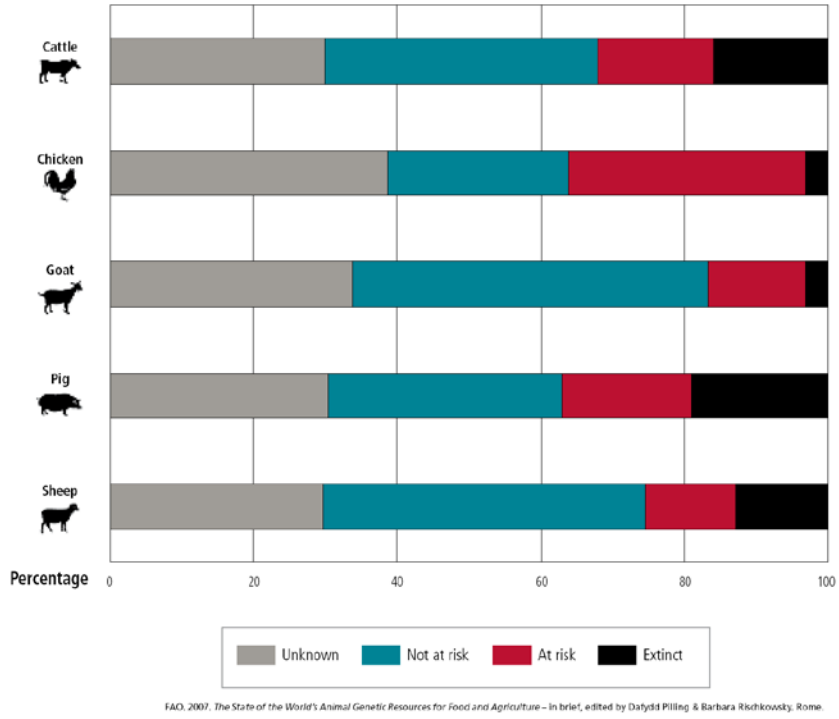

Figure 1. Breed risk status of the major livestock species in the world (FAO, 2007b).

Loss of genetic diversity can also take place within breeds in the form of loss of genes and genotypes. Strong selection in high production breeds has resulted in decreases in effective population sizes (Goddard, 1992). In many breeding programs a limited number of parents is used, resulting in loss of diversity associated with increased inbreeding. Furthermore, the less popular breeds have become smaller in population size, often resulting in loss of genetic diversity associated with increased inbreeding (Gandini and Oldenbroek, 1999).

Genetic diversity is essential for the sustainability of livestock (and other) species for a variety of reasons. First of all, genetic diversity within breeds is needed for long-term genetic improvement of livestock breeds, for selection of new traits or traits in a changing environment, and to prevent low performance due to inbreeding. Secondly, genetic diversity between breeds is important, because rare and local breeds may fulfil specific requirements that might be necessary in the future. For example, rare or local breeds may be used to support maintenance of genetic diversity in the high production breeds. Thirdly, there are historic and esthetical reasons, as many local breeds are part of our cultural heritage. And finally, many local breeds have a socio-economic value, as they can be necessary for the livelihood in harsh areas (Gandini and Villa, 2003). Therefore, it is important to prevent further loss of breeds and of diversity within breeds. The importance of conservation of genetic diversity has been recognized by many countries by signing the Convention of Biodiversity in 1992 and adopting the Global Plan of Action for Animal Genetic Resources in 2007. These initiatives highlight the responsibility and obligation of each country to conserve their native livestock breeds, and to take action to prevent further loss of genetic diversity.

1.2 Different strategies to conserve livestock genetic diversity

Conservation of genetic diversity in livestock breeds can be achieved in different ways. *In situ* conservation, defined as the conservation of live farm animals in their normal habitat, is the preferred method, because it is the most viable option in the long term. When a breed is kept in its natural environment, it can fulfill its cultural and socio-economic role, and it can adapt to changing circumstances and keep evolving. However, full attention for genetic management is needed, to prevent a breed becoming vulnerable to effects of random drift and inbreeding. If *in situ* conservation is not possible, or when the breed is threatened by genetic drift and extinction due to a small population number, *ex situ in vivo* conservation is an alternative. *Ex situ in vivo* conservation can be defined as the conservation of live animals outside their normal habitat, for example in a zoo. Because animals are

kept outside their production or natural environment, maintenance of the genetic diversity of the breed is not guaranteed. A third method is *ex situ in vitro* conservation, defined as the storage of genetic material (e.g. semen, embryo's) in liquid nitrogen. In several countries, gene banks have been set up with the aim to conserve genetic diversity between and within livestock breeds as an insurance for the future. Gene bank material can be used in different situations: to support populations to prevent or overcome genetic problems (drift, inbreeding, genetic defects), to reconstruct a breed in case of extinction or loss of a substantial number of animals, to create new lines/breeds, to quickly modify or reorient selection of a breed, or for research purposes (Hiemstra, 2003). The objectives of the Dutch gene bank and the different breeds/lines that are conserved are represented in Box 1.2. In practice, often a combination of *in situ* and *ex situ* conservation is applied, which can result in a successful conservation strategy (Oldenbroek, 2007).

Conservation of genetic diversity in a population with *in situ* or *ex situ* conservation involves two important actions. First, the amount of genetic diversity that is available in the population has to be identified. Second, animals have to be selected for conservation, both for *in situ* and *ex situ* conservation, with the main objective to conserve as much genetic diversity as possible. Such selection is necessary, because in many situations only a limited number of animals can be conserved. Estimation of genetic diversity and prioritization of animals for conservation can be done with pedigree or molecular marker information. In the next section of this introduction, first the definition of genetic diversity will be described. Then, estimation of genetic diversity with pedigree information and molecular markers will be introduced, and subsequently the conservation of genetic diversity with pedigree information and molecular markers will be discussed. Finally, we formulate the research questions and the aim of this thesis.

1.3 Definition of genetic diversity

Genetic diversity can be defined as the additive genetic variance within and between breeds or populations (Meuwissen, 2009). For genetic diversity studies it is reasonable to use the additive genetic variance, because it determines the possible response to selection. Genetic diversity in this thesis refers to the neutral additive genetic variance, which is not linked to specific traits. Genetic diversity can be estimated as the overall genetic diversity as an average measure over the whole genome, but also for one chromosome, for a chromosome region, or for smaller parts of the genome.

Box 1.2 The Dutch gene bank.
In the Dutch gene bank, a substantial semen collection has been established for most Dutch rare and several commercial domestic animal breeds (Table 1.1).

Table 1.1 Number of breeds/lines, number of donor animals and number of straws per species in the Dutch gene bank (2010).

Species	Nr of breeds/lines	Nr of animals	Nr of semen straws
Cattle	9	4,585	181,753
Dogs	2	10	162
Goats	2	30	3820
Horses	5	59	10,906
Pigs	16	519	69,981
Poultry	20	270	18,827
Sheep	7	228	23,810

The general objective of the Dutch gene bank is to conserve all rare domestic animal breeds and to stimulate animal breeders to back-up their commercial Dutch breeds or lines in the gene bank. Storage of the breeds is done using semen, but in the future it might be possible to also store other genetic material like somatic cells and embryo's. The Dutch gene bank is maintained by the Centre for Genetic Resources, The Netherlands (CGN). Activities of CGN are:

- Policy advise on conservation, management and sustainable use of animal genetic resources
- Development and management of gene bank collections of farm animals
- Research on improvement and development of methods for cryopreservation of genetic material
- Research to support conservation decisions and sustainable genetic management of breeding populations
- Monitoring of diversity in farm animals and documentation of gene bank collections and live populations
- Enhancement of international collaboration in the above areas

(Source: Brochure Maintaining the Dutch cultural heritage, www.cgn.wur.nl)

Genetic diversity within a population or breed can be estimated from the relationship between individuals, called the coancestry or the coefficient of kinship (Falconer and Mackay, 1996). The coefficient of kinship is based on the relation between kinship and diversity, where kinship is defined as the probability that two alleles drawn at random from a neutral locus are identical by descent (copies of the same ancestral allele). A high mean kinship implies low genetic diversity in the

population, which is illustrated by the relationship between kinship and additive genetic variance under Hardy-Weinberg equilibrium, $\sigma_A^2 = (1 - \bar{f})\sigma_{A,0}^2$, where $\sigma_{A,0}^2$ is the original additive genetic variance, and $\bar{f}$ the mean kinship in the current population.

1.4 Pedigree based genetic diversity estimation

Estimation of pedigree kinships with pedigree information relies on a base population: the population of animals whose parents are either unknown or ignored, and in which we define identical copies of an allele to be alike in state, but not identical by descent. In that way, pedigree kinships between animals will become higher when more generations in the pedigree become available. An average mean pedigree kinship of for example 0.25 in a population means a 25% decrease of the additive genetic variance since the base population. Pedigree kinships have been used in several studies to estimate genetic diversity within breeds (e.g. Hagger, 2005; Melka and Schenkel, 2010; Selvaggi et al., 2010). Pedigree kinship is seen as an accurate estimate for the overall loss of genetic diversity relative to the base population, provided that a reliable pedigree is available. Accuracy decreases with low pedigree depth, pedigree errors, and missing pedigree data. The estimated mean pedigree kinship in a population with a complete and deep pedigree can be very high, but when only a few generations of the pedigree are available (which means that the base population is very close to the current population), the estimated mean pedigree kinship will be much lower. Hence, this merely reflects the amount of pedigree information, rather than true differences in diversity. Furthermore, missing pedigree data and pedigree errors can result in a low estimated pedigree kinship in a population that is highly inbred, resulting in a negative effect on conservation of genetic diversity (Mucha and Windig, 2009; Oliehoek and Bijma, 2009).

1.5 Genetic diversity estimation using molecular markers

Molecular markers can be a good alternative source of information to estimate genetic diversity, in case of missing pedigree data or pedigree errors. But also when pedigree information is available, markers may allow to estimate genetic diversity more precisely, as will be explained below. An overview of the most common molecular markers that have been used for genetic diversity estimation is given in Table 1.2.

Table 1.2 Examples of use of molecular markers in genetic diversity studies.

Marker	Typical example	# markers	# alleles per marker
Blood groups	(Buys, 1990)	1	11
Allozymes	(Taggart et al., 1981)	13	2-5
AFLP	(Ajmone-Marsan et al., 2001)	219	2
RAPD	(Kantanen et al., 1995)	3-7	2
Microsatellite	(Canon et al., 2001)	16	11 (on average)
SNP	This thesis (2012)	47,213	2
Sequence	Not yet	>1,000,000	1-2

Blood groups were the first molecular markers, based on the presence or absence of inherited red cell antigens. Blood groups have been used especially in studies on cattle (Larsen and Hansen, 1986; Georges et al., 1990), but their low number prohibits genetic diversity estimation on a fine scale. Other markers used in past genetic diversity studies are allozymes, based on protein variants in enzymes. Because of their low number of loci and polymorphism level other markers have taken over (Schlotterer, 2004; Toro et al., 2009). With the arrival of new DNA techniques, nuclear DNA markers like AFLPs (amplified fragment length polymorphisms), RAPDs (randomly amplified polymorphic DNAs) and microsatellites were used in genetic diversity studies since 1990. AFLPs and RAPDs have been successfully used to analyze population genetic structures (Lynch and Milligan, 1994; Schlotterer, 2004). However, because of their dominant mode of inheritance and their difficulty to reproduce they have a reduced power to analyze within breed diversity (Schlotterer, 2004; Toro et al., 2009). Microsatellites have been the most widely used markers for genetic diversity estimation in recent years (Maudet et al., 2002; Fabuel et al., 2004; Freeman et al., 2006; Dalvit et al., 2008; Tapio et al., 2010). Microsatellites are tandemly repeated sequences, and because they are highly polymorphic and evenly distributed over the genome they have been very popular (Schlotterer, 2004).

The availability of genome-wide SNP (single nucleotide polymorphisms) markers provides new possibilities for genetic diversity estimation. A SNP marker is a single base change in a DNA sequence, with two possible nucleotides at a given position (Vignal et al., 2002). In contrast to other markers, SNP markers have a dense distribution over the genome, which enables the evaluation of genetic diversity across the whole genome in detail. SNP markers are now the markers of choice in QTL analysis and genomic selection, and already several studies used SNP data for genetic diversity estimation in livestock breeds (Zenger et al., 2007; Muir et al., 2008; Kijas et al., 2009; The Bovine HapMap Consortium, 2009; Flury et al., 2010; Lin et al., 2010; Silió et al., 2010).

Methods to estimate genetic diversity with markers are observed and expected heterozygosity (Lin et al., 2010), allelic diversity (Zenger et al., 2007), marker kinship (Eding and Meuwissen, 2001), epistatic kinship (Flury et al., 2006) and marker similarity (Lynch and Ritland, 1999). In this thesis our aim was to estimate the neutral additive genetic variance within breeds with SNP markers, for which expected heterozygosity and marker kinship were the estimates of choice. Expected heterozygosity is based on allele frequencies of SNP markers in the population, marker kinship is based on similarities between SNP markers in the population. The two estimates are directly linked to each other. In contrast to pedigree kinship, expected heterozygosity and marker kinship (both based on SNP markers) do not rely on a base population. It is merely observed whether or not markers are identical in two individuals, but a distinction between identity in state and identity by descent is not made. This means that SNP markers give a direct reflection of the genetic diversity currently present in the population, without expressing the diversity relative to a base population. However, it is possible to correct the marker kinship for the probability that markers are alike in state, as done by Eding and Meuwissen (2001). This is particularly useful with data involving multiple populations, where the lowest between-population kinship provides a natural choice for the identity in state in the base population. After this correction, marker kinships are IBD probabilities relative to the base population.

1.6 Advantages of genetic diversity estimation with SNP markers

In several situations, use of SNP markers instead of pedigree information for genetic diversity estimation can be helpful. First of all, for situations with poor or absent pedigree information. The advantage is small or absent when low density markers are used (Baumung and Solkner, 2003; Fernandez et al., 2005; Oliehoek et al., 2006), but not when large numbers of markers like SNPs (>10,000) are used. Genetic diversity can be estimated by combining pedigree and SNP data (Bömcke, 2011), or by using SNP markers only.

A second advantage is that SNP markers can be used for a more precise estimation of genetic diversity than pedigree information – even when pedigree information is available and accurate – in case the density of the SNP data is high enough. The SNP data allows to estimate the absolute genetic diversity, without relying on an arbitrary base population. Additionally, with SNP markers we can observe the Mendelian sampling, which makes it possible to observe which allele is inherited from which parent. This result in a direct reflection of the true IBD. With pedigree information

this is not possible, as pedigree based diversity only reflects the average genetic diversity. For example, full-sibs actually share between 45% and 55% of their genes rather than exactly the expected 50% (Vanraden, 2007).
A third advantage of SNP markers is that we can observe genetic diversity at specific regions over the genome. As pedigree based diversity reflects an average estimate of the genetic diversity, genome regions with higher or lower diversity cannot be identified. Identification of region-specific genetic diversity allows to identify regions with the lowest diversity, where the risk to loose genetic diversity is the greatest. When we have identified regions with low genetic diversity, we can subsequently take action to conserve the genetic diversity at these regions. How to conserve the genetic diversity at specific regions with SNP markers has yet to be investigated.

1.7 Prioritization of individuals for conservation

Prioritization has to be done in such a way that as much genetic diversity as possible is conserved, whether animals are prioritized for *in situ* or for *ex situ* conservation. Prioritization of animals is usually based on pedigree information, mainly because of the low costs and relatively simple use when pedigree information is available. The method of choice to prioritize animals with pedigree information is optimal contribution selection. Optimal contribution selection was developed to maximize genetic gain while constraining the inbreeding rate to a fixed value, but can also be used to minimize the average relatedness in the next generation. By minimizing the average relatedness among animals prioritized for conservation, the conserved genetic diversity is maximized. This method has been used in several conservation studies (Meuwissen, 1997; Grundy et al., 1998; Sonesson and Meuwissen, 2000).
Pedigree information is however not always suitable for prioritization of animals. In many situations pedigree information is unreliable or not available, and in some situations it is not possible to obtain pedigree information, like for example in wild animal populations. In that situation SNP markers are more suitable. Genotyping costs have decreased since the introduction of SNP chips, and in the future these costs will further decrease. Therefore we expect that use of SNP markers for conservation purposes will become more popular in the future. Another advantage of SNP markers for prioritization of animals is the possibility to conserve genetic diversity at specific regions or loci over the genome.
Optimal contribution selection can also be performed using SNP markers, by estimating the relatedness between animals using SNP markers. Until now, optimal contributions based on SNP markers has been mainly used to increase genetic gain while controlling inbreeding (Sonesson et al., 2010; Nielsen et al., 2011), but it may be used to prioritize animals for conservation as well.

1.8 Aim and outline of the thesis

The availability of SNP markers has resulted in new opportunities to estimate genetic diversity within livestock breeds in more detail, and to improve prioritization of animals for conservation of genetic diversity. It is hypothesized that SNP markers can give a better estimation of the genetic diversity within breeds than pedigree information, for both the overall genetic diversity and the genetic diversity at specific genome regions. We also hypothesize that SNP markers can help improve the prioritization of animals in order to conserve genetic diversity within breeds, and especially to conserve genetic diversity at specific genome regions. However, little is known about how the genetic diversity varies over the genome, and what the differences are between pedigree and SNP based diversity estimates. Additionally, we do not know how much more genetic diversity can be conserved when we use SNP markers instead of pedigree information, and what the effects are of conservation that targets a specific region or locus only. The overall objective of this thesis is to further explore the potential of SNP based genetic diversity estimators for conservation of livestock breeds.

The first aim in this thesis was to compare different methods to estimate genetic diversity with SNP markers. In order to do so, two different genetic diversity estimates based on SNP markers, expected heterozygosity and IBD probabilities, were evaluated in a simulation study (**Chapter 2**). In this study, genetic diversity at a given position on the genome was estimated by using the neighboring SNP markers. The next step was to apply one of these SNP based diversity estimates in a small Holstein cattle population (**Chapter 3**). Genetic diversity was estimated with pedigree kinship and expected heterozygosity based on SNP markers. The aim was to compare both estimates by evaluating the differences in genetic diversity for the whole genome, at the chromosomal level and at specific chromosome regions. In **Chapter 4**, the differences in prioritization of animals for conservation in a gene bank with pedigree or SNP information were investigated. This was achieved by prioritizing animals for conservation using optimal contribution selection based either on pedigree or SNP information, using two different Holstein cattle populations. The aim was to investigate the consequences of prioritization with pedigree or SNP information for the genetic diversity, by comparing the conserved genetic diversity over the whole genome and at the chromosomal level. Finally, in **Chapter 5** animals from a Holstein cattle population were prioritized for conservation in a gene bank using optimal contribution selection based on SNP information, with the focus on conserving a single locus. The aim in this study was to quantify the risk of losing

genetic diversity when conserving a single locus, and to investigate the effect of allele frequency of the single locus and population stratification on the loss of diversity.

2

Estimating genetic diversity across the neutral genome with the use of dense marker maps

Krista A. Engelsma[1,2], Mario P.L. Calus[1], Piter Bijma[2] and Jack J. Windig[1,3]

[1] Wageningen UR Livestock Research, Animal Breeding and Genomics Centre, P.O. Box 65, 8200 AB Lelystad, The Netherlands; [2] Wageningen University, Animal Breeding and Genomics Centre, P.O. Box 338, 6700 AH Wageningen, The Netherlands; [3] Centre for Genetic Resources, The Netherlands (CGN), P.O. Box 65, 8200 AB Lelystad, The Netherlands

Genetics Selection Evolution (2010) 42:12

Abstract

With the advent of high throughput DNA typing, dense marker maps have become available to investigate genetic diversity on specific regions of the genome. The aim of this paper was to compare two marker based estimates of the genetic diversity in specific genomic regions lying in between markers: IBD-based genetic diversity and heterozygosity. A computer simulated population was set up with individuals containing a single 1-Morgan chromosome and 1665 SNP markers and from this one, an additional population was produced with a lower marker density i.e. 166 SNP markers. For each marker interval based on adjacent markers, the genetic diversity was estimated either by IBD probabilities or heterozygosity. Estimates were compared to each other and to the true genetic diversity. The latter was calculated for a marker in the middle of each marker interval that was not used to estimate genetic diversity. The simulated population had an average minor allele frequency of 0.28 and an LD (r^2) of 0.26, comparable to those of real livestock populations. Genetic diversities estimated by IBD probabilities and by heterozygosity were positively correlated, and correlations with the true genetic diversity were quite similar for the simulated population with a high marker density, both for specific regions (r=0.19-0.20) and large regions (r=0.61-0.64) over the genome. For the population with a lower marker density, the correlation with the true genetic diversity turned out to be higher for the IBD-based genetic diversity. Genetic diversities of ungenotyped regions of the genome (i.e. between markers) estimated by IBD-based methods and heterozygosity give similar results for the simulated population with a high marker density. However, for a population with a lower marker density, the IBD-based method gives a better prediction, since variation and recombination between markers are missed with heterozygosity.

Key words: genetic diversity, SNP, IBD, heterozygosity, simulation, genome

2.1 Introduction

Conservation of genetic diversity in livestock is of vital importance to cope with changing environments and human demands (Oldenbroek, 2007). Intensive livestock production systems have limited the number of breeds and lines used, and many native breeds have become rare or extinct, causing a loss of genetic diversity. To conserve biodiversity and ensure its sustainable use, efforts are being made world-wide (FAO, 2007a), for example in the form of genetic diversity conservation via gene banks or by maintaining genetic diversity in breeding populations. Determining and evaluating genetic diversity present within livestock breeds are crucial to make the right conservation decisions and to efficiently use resources available for conservation.

To evaluate genetic diversity in livestock populations, several methods have been developed (Woolliams and Toro, 2007). These methods are based on pedigree information, or on molecular data when pedigree information is not available. During the last decade, availability and use of molecular information have increased, and numerous types of markers have become available to evaluate genetic diversity. Microsatellites have been widely used for conservation purposes, but are gradually being replaced by SNP markers which are available in large numbers across the entire genome. These dense marker maps enable us to evaluate genetic diversity more precisely and to obtain information on the genetic diversity separately for each specific segment of the genome.

Basically, there are two approaches to evaluate genetic diversity. In molecular and population genetics, heterozygosity of markers is the most widely used genetic diversity parameter (Toro and Caballero, 2005). In quantitative genetics and animal breeding, additive genetic variance of traits estimated with the help of pedigrees is generally used to evaluate genetic diversity (Falconer and Mackay, 1996). To determine additive variance with markers, the probability that two alleles are identical by descent (IBD), i.e. originate from the same ancestral genome, is estimated (Meuwissen and Goddard, 2001). The probability of IBD is closely related to the relationship coefficient (r) calculated from pedigrees for the estimation of additive variance. Although theoretically both approaches should give similar results, in practice they are weakly correlated (Reed and Frankham, 2001; Toro et al., 2009). As dense marker maps have become available, it is possible to estimate additive genetic effects of markers and this is routinely used in, for example, QTL-detection (Fernando and Grossman, 1989) and genomic selection (Meuwissen et al., 2001; Calus et al., 2008) .

A crucial difference between heterozygosity on the one hand and IBD probabilities and *r* on the other hand is that the latter depend on a base population. Markers can be alike in state (AIS) but not IBD if they originate from different ancestors in the base population. With heterozygosity this distinction is not made. For example, in the case of QTL detection, IBD probabilities are used because they better predict whether two chromosome intervals carry the same QTL. The reason is that if an individual carries markers at two loci around an interval that are both AIS, but not IBD (i.e. originate from different ancestors), it is less likely that the interval between the markers is completely AIS and carries the same QTL. However, if both markers are IBD the interval will also be IBD (and AIS), unless a double recombination has occurred in the interval.

Both heterozygosity and IBD probabilities can be used to estimate genetic diversity in specific regions of the genome, in which it may deviate from the average diversity calculated over the whole genome. Heterozygosity and IBD probabilities as genetic diversity measures may also deviate from each other. It is unclear how substantial the difference is between the two approaches and whether it varies over the genome. These local differences may be averaged out if the average diversity is calculated over the whole genome. However, both approaches can be used to estimate the genetic diversity for sequences lying in between genetic markers. Because IBD probabilities are used specifically to predict the presence of QTL between markers one may expect that IBD probabilities better predict genetic variation between markers. Whether this is a substantial difference is not clear.

The aim of this paper was to compare two different estimates of the genetic diversity of a region lying in between markers over the genome i.e. IBD probabilities between marker haplotypes and heterozygosity. Towards this aim, we generated genetic diversity over a genome by computer simulation of two populations each with a different marker density. IBD-based genetic diversity and heterozygosity were compared for the average diversity of regions in the genome containing several marker intervals, and for the genetic diversity at each marker interval. To evaluate how well these estimates predict the genetic diversity over the genome, both were compared to the true genetic diversity.

2.2 Material and methods

A population was computer simulated with neutral SNP markers across the genome. Next, for each locus in the genome, the genetic diversity was estimated in three ways: (1) based on IBD probabilities with flanking markers; (2) based on expected heterozygosity with flanking markers; (3) the true expected heterozygosity of the

marker itself. For (1) and (2), the marker at the locus itself was assumed to be unknown. In this way the predicted diversities (1) and (2) could be compared with true genetic diversity (3).

Simulated population

Simulations were aimed at generating a population with a neutral genetic diversity varying over the genome. We avoided selection as this may cause specific patterns in genetic diversity (e.g. selective sweeps). Variation in diversity in the simulated population was generated by random mating, recombination, mutation and sampling of maternal and paternal chromosomes. The simulated population started with 1000 animals with an equal sex ratio, and this structure was kept constant for 1000 generations. Animals were mated by drawing parents randomly from the previous generation, and mating resulted in 1000 offspring (500 males and 500 females) in each generation. A genome containing a single 1-M chromosome was simulated, starting with 2000 SNP marker loci with positions on the genome determined at random. This density is roughly equivalent to the current SNP chips available for livestock species (e.g. 50K SNP chip for the 30-M genome in cattle). In the first generation (base population), marker loci were coded as 1 or 2 and allocated at random, so that allele frequencies (p) averaged 0.5. This was comparable to the simulation used in the study of Habier et al. (2007). During the simulation of the 1000 generations, marker alleles were dispersed through the population by random mating, recombinations and mutations. Recombinations between adjacent loci occurred with a probability calculated with Haldane's mapping function, based on the distance between the loci. Mutations occurred for each locus only once during the 1000 generations, where mutations changed the allele state from 1 to 2 or from 2 to 1, with equal probability. Three additional generations were simulated after the first 1000 generations, which were assumed to be genotyped, to analyze genetic diversity over the genome, e.g. similarly as in livestock breeds where only recent generations are genotyped. All SNP markers with a minor allele frequency in generations 1002 and 1003 of <0.02 were discarded from the analysis. Thus, the generated population consisted of 3000 animals (generation 1001, 1002 and 1003) with a known genotype, and 1665 SNP markers were still segregating in these generations.

To determine whether marker density would influence the genetic diversity estimation with the different estimates, a second population was obtained with a lower marker density. This population was based on the first population, by changing only the number of SNP markers from 1665 to 166, by systematically deleting 90% of the SNP markers.

IBD probabilities

Genetic diversity was estimated for each marker interval on the genome. A marker interval was defined as the interval between two genotyped markers, with one marker lying in between these two markers which was not taken into account for the genetic diversity estimation (ungenotyped marker) (Figure 2.1). In the next marker interval, this middle ungenotyped marker became the flanking marker of the interval with the adjacent marker being the ungenotyped marker. The genetic diversity estimation was based on IBD probabilities between haplotypes, where a haplotype was defined as a combination of ten consecutive markers, i.e. five markers on either side of the marker interval (Meuwissen and Goddard, 2001). Haplotypes were reconstructed from the genotypes using the methods of Windig and Meuwissen (2004). By using IBD probabilities, the chance of markers being similar (AIS) but not IBD is taken into account. This contrasts with heterozygosity, where similar markers are all assumed to originate from the same ancestor (AIS=IBD). Additionally, because haplotypes were used, the recombination history is taken into account to estimate the probability of IBD. For example, a long string of identical markers strongly indicates a recent common ancestor (probability of being IBD must be high), because strings of identical markers from non-recent ancestors are generally broken up by recombination.

IBD probabilities were calculated between the existing haplotypes in the simulated population for each marker interval, by combining linkage disequilibrium and linkage analysis information, where both pedigree and marker information were used. IBD probabilities were first calculated for the first generation of genotyped animals, using the algorithm of Meuwissen and Goddard (2001). In this method, IBD probabilities are calculated for a fictitious locus A in the middle of a marker interval, where information is used from the markers on either side of this locus A.

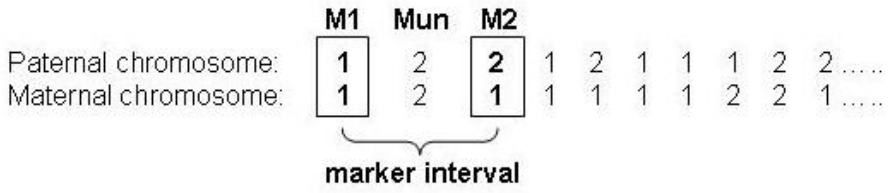

Figure 2.1 Definition of marker interval, ungenotyped marker (Mun), and adjacent markers (M1, M2, …) used for the genetic diversity estimation. The ungenotyped marker is placed in the middle of the marker interval. Genetic diversity was estimated for each marker interval, using the adjacent markers left and right of the interval.

In our case, locus A is positioned at the marker locus in the middle of each marker interval. The probability of A in two haplotypes being IBD or not IBD is estimated by weighing all possible combinations of the markers in the haplotype being IBD or not IBD with recombinations. The IBD probability is calculated back to an arbitrary base population, T generations ago (we used T=1000). In this calculation, effective population size (we used Ne=1000 during the 1000 generations) and recombination probabilities based on marker distances are taken into account. As the number of markers with identical alleles increases, the probability that the two fictitious alleles for A are IBD also increases.
After calculating IBD probabilities for the haplotypes in the base generation, the haplotypes of the animals in later generations were added, and the elements in the IBD matrix for those descendant haplotypes were calculated using the algorithm of Fernando and Grossman (1989). In this algorithm, IBD probabilities between offspring are calculated based on the IBD probabilities between the parents and the inheritance of the markers (Meuwissen and Goddard, 2001). Whenever the IBD probability of descendant haplotypes with one of their parental haplotypes exceeded 0.95, the descendant haplotype was clustered with this parental haplotype. This was done to avoid excessive numbers of near identical haplotypes resulting in long computation times.

Genetic diversity based on IBD probabilities

The genetic diversity for all marker intervals on the genome in the simulated population was estimated using haplotype frequencies and IBD probabilities between haplotypes. Haplotype frequencies (frequency of the different haplotype configurations in the population) per marker interval were obtained by:

$$\boldsymbol{c}_i = N_{ij} / N_i \qquad (1)$$

where $\boldsymbol{c}_i$ is a contribution vector with haplotype frequencies for all haplotypes on marker interval i, N_{ij} is the number of haplotypes of type j on marker interval i, and N_i is the total number of haplotypes in the population on marker interval i.
Genetic diversity per marker interval was determined by calculating the average haplotype relatedness at each locus (Meuwissen, 1997):

$$r_i = \boldsymbol{c}_i' \boldsymbol{IBD}_i \boldsymbol{c}_i \qquad (2)$$

where r_i is the average relatedness for marker interval i, and $\boldsymbol{IBD}_i$ is the IBD-matrix for marker interval i. The genetic diversity for marker interval i was calculated as:

$$GD\_IBD_i = 1 - r_i \quad (3)$$

This is the predicted probability that the marker in the middle of the interval is not IBD.

Heterozygosity

Expected heterozygosity (Falconer and Mackay, 1996) was calculated for each marker interval on the genome in the simulated population, using one flanking marker on either side of the interval. Heterozygosity was calculated in two different ways: average heterozygosity of the two adjacent markers around the marker interval (H_{exp}_AVG), and heterozygosity for the interval treating both markers as a single two-marker haplotype (H_{exp}_HAP2). For the calculation of H_{exp}_AVG, first expected heterozygosity was calculated for the markers on the left and right of the interval separately (see Figure 2.1, markers on the left and right of the interval are in bold):

$$H_{\exp,j} = 2p_j q_j \quad (4)$$

where *p* and *q* are the allele frequencies for marker *j* in the simulated population. Subsequently, the expected heterozygosity for each marker interval (H_{exp}_AVG) was calculated by taking the average of the expected heterozygosity for both markers left and right of the marker interval.

H_{exp}_HAP2 was calculated for the combination of the two markers on the left and right of the interval as a two-marker haplotype (see Figure 2.1, haplotype is shown with the two markers in bold), where four combinations were possible (11, 12, 21, and 22). H_{exp}_HAP2 for marker interval *i* was calculated as:

$$H_{\exp}\_HAP2_i = 1 - \sum_k p_i^2 \quad (5)$$

where p_i is the frequency of the haplotype with combination *k* at marker interval *i*.

Comparison GD_IBD and heterozygosity

Comparison between genetic diversity measures GD_IBD, H_{exp}_AVG and H_{exp}_HAP2 was done by calculating Pearson's correlations. Correlations were calculated between the genetic diversity measures for each marker interval, but also between

the measures averaged over groups of adjacent marker intervals, to investigate whether the correlations would change when the measures were averaged over larger regions of the genome. Therefore, correlations were calculated between GD_IBD, H_{exp}_AVG and H_{exp}_HAP2 for 4, 10, 20 and 40 marker intervals together. For example, for 10 marker intervals together, the correlations were calculated with the average measures for interval 1-10, 11-20, 21-30, etc.

Comparison with true diversity

To evaluate whether one of the approaches better predicts genetic diversity, a true genetic diversity was calculated for the ungenotyped marker lying within each marker interval. This marker was not used to estimate genetic diversity with GD_IBD, H_{exp}_AVG and H_{exp}_HAP2, but the adjacent markers were used to predict the diversity in this ungenotyped marker. The true genetic diversity for the ungenotyped marker in the marker interval was determined by calculating the expected heterozygosity (Equation 4). To compare true genetic diversity (H_{exp}_TRUE) with GD_IBD and heterozygosity (H_{exp}_AVG and H_{exp}_HAP2), Pearson's correlations were calculated for each marker interval and for groups of marker intervals (4, 10, 20 and 40). Two correlations were estimated for each comparison: between true genetic diversity of the even markers and their estimated genetic diversity based on the uneven (flanking) markers, and the other way around. This was done because the genotyped marker in one marker interval became the ungenotyped marker in the next marker interval.

2.3 Results

Simulated population

In the simulated data, 1665 SNP markers were still segregating in generations 1001, 1002 and 1003. Marker distances ranged from 0.00 cM to 0.50 cM, with an average of 0.06 cM. The number of marker haplotypes used for GD_IBD after clustering varied from 1 to 56, with an average of 20.70 haplotypes. The average minor allele frequency over the 1665 SNP markers was 28%, ranging from 2 to 50%. The average linkage disequilibrium (r^2) between adjacent markers, calculated as the square of the correlation of allele frequencies (Hill and Robertson, 1968), was 0.26. The simulated population was comparable to real livestock populations. For example, in cattle nowadays ~50,000 SNPs are used for a 30-M genome, which gives an average marker distance of 0.06 cM. On the cattle 50k SNP chip, for HF dairy cattle the r^2 between adjacent markers is between 0.15 and 0.20 for an average marker distance of ~0.06 cM (De Roos et al., 2008; Khatkar et al., 2008).

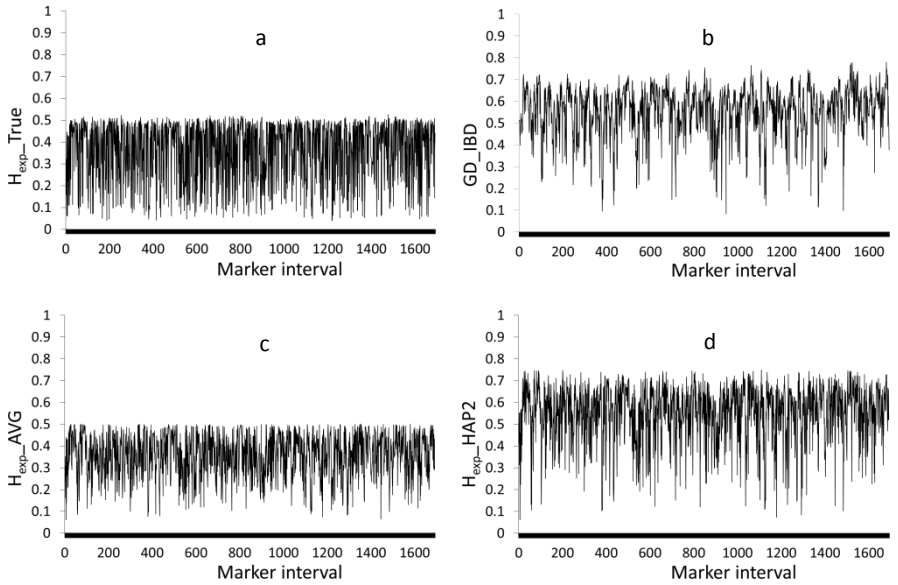

Figure 2.2abcd Distribution of the estimated genetic diversity across the simulated genome. (a) True genetic diversity calculated by expected heterozygosity for the ungenotyped marker loci within the marker interval (H_{exp}_TRUE); (b) Estimated genetic diversity with IBD probabilities between marker haplotypes (GD_IBD); (c) Estimated genetic diversity with expected heterozygosity as an average for the two flanking markers (H_{exp}_AVG); (d) Estimated genetic diversity with expected heterozygosity for the two flanking markers as a two marker haplotype (H_{exp}_HAP2).

The true genetic diversity over the simulated genome, calculated as the expected heterozygosity for the marker locus within each marker interval (Hexp_TRUE), ranged from 0.04 to 0.53 with an average of 0.36 (Figure 2.2a). A large number of H_{exp}_TRUE values was found between 0.48 and 0.50 (Figure 2.3a), which is in accordance with a population in Hardy-Weinberg equilibrium for an allele frequency range 0.4-0.5.

Genetic diversity estimates

Genetic diversity estimated by IBD probabilities (GD_IBD) varied considerably over the genome, with values ranging from 0.00 to 0.75, with an average of 0.52 (Figures 2.2b and 2.3b). Expected heterozygosity calculated for the two adjacent marker loci around each marker interval as an average (H_{exp}_AVG) resulted in systematically lower values with a smaller range compared to GD_IBD (0.05 to 0.50, average of 0.36) (Figures 2.2c and 2.3c). When expected heterozygosity was calculated for flanking markers as a two-marker haplotype (H_{exp}_HAP2), the level

and range of values increased and were more similar to GD_IBD (0.05 to 0.75, average of 0.55) (Figures 2.2d and 2.3d). This result was expected, since genetic diversity estimation with H_{exp}_HAP2 is more similar to GD_IBD because H_{exp}_HAP2 also uses a haplotype construction, but with only two markers instead of ten. Both heterozygosity estimates fluctuated more over the genome compared to GD_IBD, reflecting a lower correlation between values of adjacent marker intervals for the heterozygosity estimates (H_{exp}_AVG: r=0.23; H_{exp}_HAP2: r=0.28; GD_IBD: r=0.64).

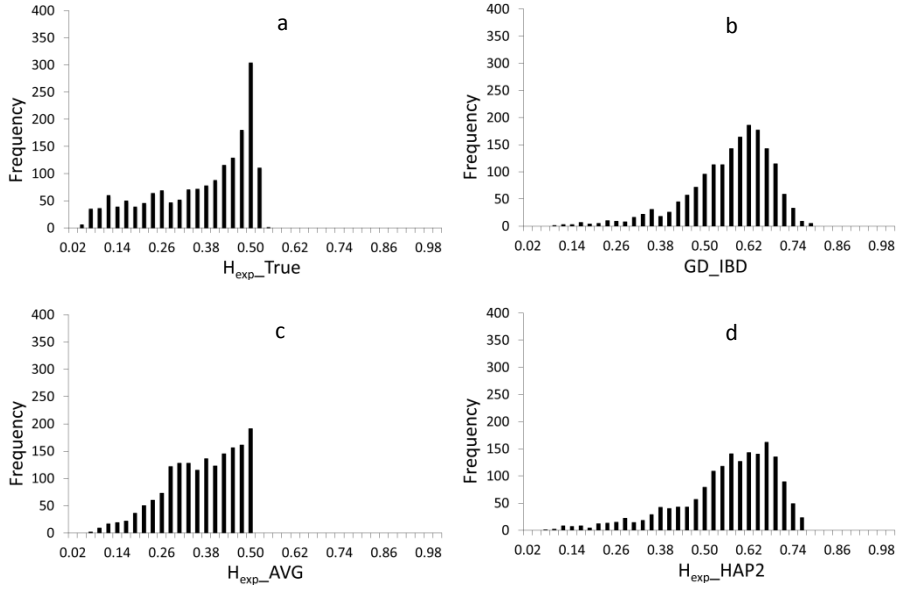

Figure 2.3abcd Frequency of the estimated genetic diversity across the simulated genome. (a) True genetic diversity calculated by expected heterozygosity for the ungenotyped marker loci within the marker interval (H_{exp}_TRUE); (b) Estimated genetic diversity with IBD probabilities between marker haplotypes (GD_IBD); (c) Estimated genetic diversity with expected heterozygosity as an average for the two flanking markers (H_{exp}_AVG); (d) Estimated genetic diversity with expected heterozygosity for the two flanking markers as a two marker haplotype (H_{exp}_HAP2).

Comparison with true genetic diversity

The correlation between H_{exp}_TRUE and GD_IBD was weak (r=0.21), and comparable to the correlations between H_{exp}_TRUE and H_{exp}_AVG (r=0.19) and H_{exp}_HAP2 (r=0.20) (Table 2.1 and Figure 2.4). These results indicate that both GD_IBD and heterozygosity estimates are similar in predicting the genetic diversity for ungenotyped regions of the genome in the current simulated population. The correlation between GD_IBD and H_{exp}_AVG was 0.46, and was slightly higher between GD_IBD and H_{exp}_HAP2 (r=0.49) (Table 2.1).

Comparison with true genetic diversity averaged over marker intervals

When GD_IBD, H_{exp}_AVG and H_{exp}_HAP2 were averaged over groups of marker intervals, the correlations between H_{exp}_TRUE and these estimates increased. They were moderate when estimates were averaged over 40 marker intervals (r=0.61-0.64, Table 2.1). Correlations of all three estimates with H_{exp}_TRUE were comparable to each other. The correlation between GD_IBD and heterozygosity estimates H_{exp}_AVG and H_{exp}_HAP2 increased with an increasing number of marker intervals, and in the case of 40 marker intervals equaled 0.75 and 0.82, respectively. This indicates that GD_IBD, H_{exp}_AVG and H_{exp}_HAP2 are similar in predicting the genetic diversity for specific regions of the genome in a population with a high marker density.

Table 2.1 Correlations of true genetic diversity (H_{exp}_TRUE) with IBD-based diversity (GD_IBD) and heterozygosity (H_{exp}_AVG and H_{exp}_HAP2).

MI [a]	True vs. GD_IBD [b]	True vs. H_{exp}_AVG [b]	True vs. H_{exp}_HAP2 [b]	GD_IBD vs. H_{exp}_AVG [b]	GD_IBD vs. H_{exp}_HAP2 [b]
1	0.20	0.19	0.20	0.46	0.49
4	0.33	0.27	0.28	0.54	0.58
10	0.46	0.37	0.38	0.64	0.70
20	0.56	0.47	0.50	0.73	0.80
40	0.62	0.61	0.64	0.75	0.82

[a] The number of marker intervals taken into account to estimate the genetic diversity.

[b] Correlations were calculated for values per marker interval, and for average values for a group of marker intervals (4, 10, 20 and 40 marker intervals); for the latter, correlations were calculated for the true genetic diversity of even ungenotyped markers with the estimated genetic diversity based on uneven (flanking) markers, and the other way around; the average of both correlations (even and uneven) is presented.

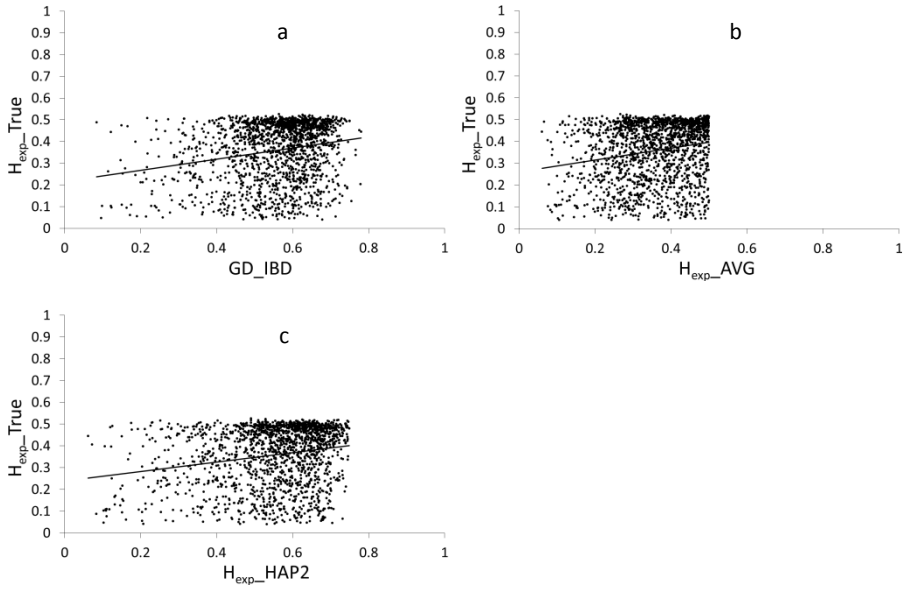

Figure 2.4abc Relationship between the true genetic diversity (H_{exp}_TRUE) and estimated genetic diversities. (a) by IBD probabilities between marker haplotypes (GD_IBD); (b) by expected heterozygosity as an average for the two flanking markers (H_{exp}_AVG); (c) by expected heterozygosity for the two flanking markers as a two marker haplotype (H_{exp}_HAP2).

Influence of marker density

When genetic diversity over the genome was estimated in a population with a lower marker density, the correlations between the true genetic diversity and GD_IBD, H_{exp}_AVG and H_{exp}_HAP2 changed, and turned out to be slightly higher for GD_IBD (Table 2.2). This result suggests that GD_IBD is a better predictor for genetic diversity when using marker maps with a lower marker density.

Table 2.2 Correlations of true genetic diversity (H_{exp}_TRUE) with IBD-based diversity (GD_IBD) and heterozygosity (H_{exp}_AVG and H_{exp}_HAP2), for a low marker density population (166 SNPs).

MI [a]	True vs. GD_IBD [b]	True vs. H_{exp}_AVG [b]	True vs. H_{exp}_HAP2 [b]	GD_IBD vs. H_{exp}_AVG [b]	GD_IBD vs. H_{exp}_HAP2 [b]
1	0.15	0.06	0.04	0.43	0.43
4	0.34	0.18	0.20	0.53	0.53
10	0.51	0.41	0.46	0.79	0.77
20	- [c]	- [c]	- [c]	- [c]	- [c]
40	- [c]	- [c]	- [c]	- [c]	- [c]

[a] The number of marker intervals taken into account to estimate the genetic diversity.
[b] Correlations were calculated for values per marker interval, and for average values for a group of marker intervals (4 and 10 marker intervals); for the latter, correlations were calculated for the true genetic diversity of even ungenotyped markers with estimated genetic diversity based on uneven (flanking) markers, and the other way around; the average of both correlations (even and uneven) is presented.
[c] There were not enough estimates left over to calculate the correlation.

2.4 Discussion

The aim of this paper was to compare two different estimates of genetic diversity of a region lying in between markers over the genome i.e. IBD-based genetic diversity and heterozygosity. Genetic diversities estimated by IBD probabilities and by heterozygosity of flanking markers were positively correlated. The correlation of GD_IBD and heterozygosity with the true genetic diversity was quite similar for a simulated population with a high marker density, for both specific and large regions over the genome. For a population with a lower marker density, GD_IBD turned out to be a better predictor of genetic diversity.

The assumption that is made for genetic diversity in the ungenotyped marker interval is different for GD_IBD and heterozygosity. With GD_IBD the assumption is that in the base population relatedness was 0, i.e. all markers were not-IBD and "heterozygosity" was 100%. With heterozygosity, no such base population is assumed and the assumption is that heterozygosity in the current generation for genotyped markers is predictive for ungenotyped markers. This explains why the average GD_IBD estimated in this study was higher than the heterozygosity estimates and the true heterozygosity. Heterozygosity based on SNP markers with only two alleles will have, under HWE, a maximum heterozygosity of 50% when the minor allele frequency is 50%, as was simulated in this study. For markers that have an unlimited number of alleles, the true heterozygosity would probably be on average closer to GD_IBD, while for markers with a low diversity the true heterozygosity would be below both GD_IBD and heterozygosity estimates.

When the genotyped marker is actually part of the gene of interest, e.g., when the marker is a known QTL, then heterozygosity at the marker fully determines the additive genetic variance due to the QTL. In that case, additive genetic variance due to the QTL simply equals $H_{exp}\alpha^2$, α denoting the allele substitution effect of the gene (Falconer and Mackay, 1996). Hence, when markers coincide with genes of interest, i.e. there are no QTL other than the genotyped markers, there is no need to consider IBD probabilities. However, in most cases, the genes of interest and their QTL will be unknown, and it is unlikely that they coincide precisely with genotyped markers. Consequently, prediction of diversity in the ungenotyped regions between markers is more relevant than the expected diversity at the markers, because most genes of interest will be in the regions between two markers. Such a prediction requires LD between the genotyped markers and the regions in-between markers, similar to the requirements in QTL mapping (Dekkers and Hospital, 2002). Our results show that the IBD-based method and heterozygosity are similar in using LD information in the current simulated data with 1665 SNP markers. However, when a population with a lower marker density was used, GD_IBD became a slightly better predictor of the genetic diversity in the marker interval. In this second population the LD between markers is low due to a larger marker distance, and in that case the IBD-based method was expected to be a better predictor, based on QTL mapping and genomic selection studies. Explaining genetic diversity at a ungenotyped locus is similar to the approaches of QTL mapping and genomic selection, where the objective is to predict genetic variance at one or more unobserved QTL. In those approaches, it has been shown that using an IBD-based method to predict genetic variance at the unobserved QTL is beneficial when the LD between the marker(s) and the QTL is low, while this benefit disappears when the LD increases (Grapes et al., 2004; Calus et al., 2008).

In our study we ignored the non-segregating SNP markers, as these markers are fixed in the simulated population and show no variation. This can be compared with common practice where base pairs for which no SNP markers are detected are considered uninformative. However, we do not know whether this variation was never there or existed in earlier generations and disappeared. In the latter case, these base pairs indicate a genetic diversity of 0, and should not be ignored. In addition, when non-segregating markers are used in another population, they might show variation and become informative. However, the correlations between the different estimates for genetic diversity as estimated in this paper are unlikely to be influenced by the exclusion of non-segregating markers.

In this study, the estimation of genetic diversity was done for a neutral genome without selection. The correlation between genetic diversity estimates and true genetic diversity was weak, but might increase if adaptive trait variation is taken into account. The availability of dense marker maps has opened up new possibilities to identify reduced or increased levels of variability on specific regions of the genome, associated to functional genes (Toro et al., 2009). In case of selection, larger regions with less variation can be found on the genome (Toro and Maki-Tanila, 2007) and a better prediction of the genetic diversity is possible.

How well the two methods predict genetic diversity depends on the variation in diversity between adjacent markers. In contrast to GD_IBD, the heterozygosity estimates assume that diversity is similar for adjacent markers and for instance ignore recombination. When regions of the genome form 'haplotype blocks', adjacent markers have (near) identical diversity. In this case, heterozygosity will better predict the genetic diversity. This was seen when we simulated a population with an effective population size of 100 instead of 1000, and 'haplotype blocks' occurred due to the loss of variation. In this population the correlation between the heterozygosity estimate H_{exp}_AVG and the true genetic diversity was higher compared to the correlation between GD_IBD and the true genetic diversity (0.97 and 0.90, respectively). However, when a population contains more variation, diversity in between markers can be missed by heterozygosity, as heterozygosity is only based on the variation of the markers itself. In that situation, GD_IBD also takes into account the variation and possible recombination in between markers, and is then expected to be a better estimator of the genetic diversity over the genome. Consequently, as shown in this study the method of choice will also depend on the marker density (Grapes et al., 2004; Calus et al., 2008), with high marker densities (i.e. > 50 markers per cM) heterozygosity is likely to perform better, with lower marker densities (i.e. <10 markers per cM) GD_IBD is likely to perform better.

2.5 Conclusions

In conclusion, the IBD-based method and heterozygosity used to estimate genetic diversity of ungenotyped regions of the genome (i.e. between markers) give similar results for a simulated population with a high marker density. However, for a population with a lower marker density, the IBD-based method gives a better prediction, since variation and recombination between markers are missed with heterozygosity. IBD-based methods can provide more insight in the genetic diversity of specific regions of the genome, and subsequently contribute to select

more accurately the animals to be conserved, for example, to construct a gene bank.

2.6 Acknowledgements

This study was financially supported by the Ministry of Agriculture, Nature and Food (Program “Kennisbasis Dier”, code: KB-04-002-021). The authors would like to acknowledge Sipke Joost Hiemstra and Johan Van Arendonk for their advice on the research and first draft of the paper, and Han Mulder for his assistance in the analysis.

3

Pedigree- and marker-based methods in the estimation of genetic diversity in small groups of Holstein cattle

K.A. Engelsma[1,3], R.F. Veerkamp[1], M.P.L. Calus[1], P. Bijma[2] and J.J. Windig[1,3]

[1] Wageningen UR Livestock Research, Animal Breeding and Genomics Centre, P.O. Box 65, 8200 AB Lelystad, The Netherlands; [2] Wageningen University, Animal Breeding and Genomics Centre, P.O. Box 338, 6700 AH Wageningen, The Netherlands; [3] Centre for Genetic Resources, The Netherlands (CGN), P.O. Box 65, 8200 AB Lelystad, The Netherlands

Journal of Animal Breeding and Genetics (2012) 129:3, 195-205

Abstract

Genetic diversity is often evaluated using pedigree information. Currently, diversity can be evaluated in more detail over the genome based on large numbers of SNP markers. Pedigree- and SNP-based diversity were compared for two small related groups of Holstein animals genotyped with the 50k SNP chip, genome-wide, per chromosome and for part of the genome examined. Diversity was estimated with coefficient of kinship (pedigree) and expected heterozygosity (SNP). SNP-based diversity at chromosome regions was determined using 5-Mb sliding windows, and significance of difference between groups was determined by bootstrapping. Both pedigree- and SNP-based diversity indicated more diversity in one of the groups; 26 of the 30 chromosomes showed significantly more diversity for the same group, as did 25.9% of the chromosome regions. Even in small populations that are genetically close, differences in diversity can be detected. Pedigree- and SNP-based diversity give comparable differences, but SNP-based diversity shows on which chromosome regions these differences are based. For maintaining diversity in a gene bank, SNP-based diversity gives a more detailed picture than pedigree-based diversity.

Key words: diversity, gene banks, pedigree, SNP

3.1 Introduction

Maintaining genetic diversity in livestock breeds has become even more important since the globalization of breeding programs (FAO, 2009). Over the last decades, genetic diversity of livestock populations had been alternatively measured using pedigree information or microsatellite data when genealogy is not available. Currently, the availability of high-density SNP chips has opened up new opportunities to evaluate genetic diversity based on genetic markers. Up to now, conservation decisions for gene banks were often based on pedigree information, while the use of high-dense markers may give a more detailed picture of the diversity across the genome.

The correlation between pedigree-based diversity and molecular diversity depends on the number of markers, their frequencies and on the dispersion of the coancestry coefficient. The correlation is weak when a few markers are used, because the markers only reflect inbreeding at some (random) points along the genome, while pedigree-based diversity gives an overall estimate. For example, in humans, the correlation between pedigree inbreeding and homozygosity based on 410 microsatellite markers and 10,000 SNPs was 0.39 and 0.56, respectively (Carothers et al., 2006), and in Holstein sires, the correlation between pedigree inbreeding and multilocus homozygosity with 10,000 SNPs was around 0.5 (Daetwyler et al., 2006). With larger number of markers, stronger correlations are possible; for example, in an Iberian pig population genotyped for 60,000 SNPs, this correlation was much higher (0.92) (Silió et al., 2010). Correlations of unity are unlikely reached because Mendelian sampling is ignored in pedigree-based inbreeding. Moreover, pedigree information can be incomplete or wrong.

In genetic conservation of livestock, diversity of often small groups of related animals within a breed has to be compared. Examples are when herds have to be prioritized for support or when the amount of diversity conserved in a gene bank is evaluated. In endangered breeds, often only a few animals remain. In European cattle, there are, for example, 13 breeds with < 100 females (Duclos and Hiemstra, 2010). Typically, differences in diversity are small in such groups because all animals tend to be related. Under these conditions, the effect of Mendelian sampling can be pronounced. Consequently, using dense SNP marker maps may be an attractive alternative to pedigree-based diversity.

Another source of differences between pedigree- and marker-based diversity is the difference in definition of the founder population: identical by descent (IBD) versus identical by state (IBS). Pedigree-based diversity is estimated as the probability that two alleles drawn randomly from two individuals are IBD, indicating that they

descend from the same ancestor since the base population (Falconer and Mackay, 1996). This base population is the founder population in the pedigree, in which all alleles are defined as being not IBD. Alleles being identical in the base population are IBS but not IBD. With marker-based diversity, the probability of alleles being IBD is estimated without reference to a base population, and therefore, all alleles being IBS are assumed to be IBD (Oliehoek et al., 2006; Powell et al., 2010). Consequently, pedigree- and marker-based diversity is estimated on a different scale, as pedigree-based diversity reflects only diversity as a result of recent ancestry.

Large numbers of SNPs have been used in Holstein cattle to estimate effective population size and divergence between different populations (De Roos et al., 2008; Flury et al., 2010), but never to evaluate diversity for conservation purposes. Different animals may be prioritized for inclusion in a gene bank when the prioritization is based on SNP-estimated diversity instead of pedigree-based diversity (Engelsma et al., 2011). A next important step is to investigate whether diversity of groups of animals within a breed is estimated differently, using pedigree or SNP information, and whether there is variation in diversity across the genome.

The objective of this study was to compare diversity based on pedigree information with diversity based on SNP information over the whole genome, at the chromosomal level and at specific chromosome regions. Specifically, we want to determine whether the difference in diversity between two small groups of related animals depends on the type of analysis (SNP-based versus pedigree-based) and on the part of the genome examined. For this purpose, two groups of Holstein animals were available that were genotyped with the 50 k SNP chip.

3.2 Material and methods

Animals

To compare pedigree- and SNP-based diversity, 90 Holstein Friesian heifers were used, consisting of two groups of animals. Although somewhat arbitrarily, the two groups might reflect different herds, bloodlines or flocks that are typically present in small populations that need to be considered for gene banks. For another experiment, the groups were selected for a high (51 heifers) or low (39 heifers) genetic production index for milk, fat and protein (Inet) and came from the same population. Selection was based on the pedigree index of the sire and maternal grandsire. Heifers were purchased in 2003 from 61 different farms throughout the Netherlands. The difference in Inet between the two groups was on average 195 Euros, representing 10 years of ongoing selection at that time (Beerda et al., 2007).

Heifers were 100% Holstein Friesian (n=84) or 87.5% Holstein Friesian and 12.5% Dutch Friesian (n=6, of which 1 with high and 5 with low production). Although differences in terms of generations selection were small, substantial differences in phenotypes were observed (Beerda et al., 2007; Windig et al., 2008).
Next to the selection differences, the two groups differed slightly in their pedigree completeness. The average number of discrete generation equivalents (Woolliams and Mantysaari, 1995) was eight generations for all 90 heifers, with 7.6 generations (ranging from 6.3 to 8.6) in the group with high breeding values (further on group EBV_{high}) and 8.1 (ranging from 6.3 to 8.9) for the group with low breeding values (further on group EBV_{low}) (Table 3.1). The heifers were sired by 49 different bulls, with 23 sires in group EBV_{high} and 28 sires in group EBV_{low}, with on average 2.2 and a maximum of six offspring per sire for group EBV_{high}, and on average 1.4 and a maximum of four offspring per sire for group EBV_{low} (Table 3.1). Based on these results, we expected the relatedness in group EBV_{high} to be somewhat higher.

Table 3.1. Pedigree information for each Holstein group (high and low EBV).

	Group EBV_{high}	Group EBV_{low}
Number of individuals	51	39
Contribution of Dutch Friesian (%)	0.25	2.88
Average number of known generations	7.6	8.1
Number of sires	23	28
Average number of offspring per sire	2.2	1.4
Maximum number of offspring per sire	6	4
Average birth date sires	19-07-1994	19-01-1995
Average inbreeding coefficient	0.057	0.036

Genotyping

For the evaluation of the genetic diversity, DNA was extracted from 96 heifers and used to determine genotypes at 54,001 SNP loci with the Illumina Bovine SNP50 Bead Chip (Illumina Inc., San Diego, CA) array. SNP quality was checked before analysis, and for this check, we used an additional dataset of 600 genotyped Holstein cows genotyped at the same time. This group of 600 animals consisted of a mixture of animals of various origin used in different experiments and could not be used for the comparison of genetic diversity of different groups. First of all, animals with >5% missing SNP genotypes were removed. SNPs without known position on the genome were removed from the dataset, and for each SNP to be included in the data, we used a call rate of more than 90%, a GenCall score more than 0.2 and a GenTrain score more than 0.55 (Illumina descriptive statistics relating to genotype quality). Additionally, SNPs with extreme deviations from

Hardy-Weinberg equilibrium were removed (chi-square test $\chi^2 > 600$, following Wiggans et al. (2009)). Non-segregating SNPs, and SNPs with a very low minor allele frequency (MAF), were not removed from the dataset as we were also interested in the regions on the genome with extremely low or no variation. After all editing steps, 90 animals and 47,213 SNPs were left and used in the analysis. The number of SNPs is higher than what generally remains after the editing steps in, for example, genome association studies (Schulman et al., 2011), because SNPs with low MAF and SNPs from the X chromosome are included.
Remaining SNPs were phased using the software package fastPHASE (Scheet and Stephens, 2006). This implied attributing alleles to one of the chromosomes (paternal or maternal) and imputing missing alleles, based on haplotype frequencies in the population. Phasing was carried out for the whole population of 690 animals. The percentage of missing alleles before imputation was with 0.10% for group EBV_{high} (ranging from 0 to 0.72%) and 0.06% for group EBV_{low} (ranging from 0.01 to 0.2%) very low.

Pedigree-based diversity

The overall genetic diversity based on pedigree information in the two groups was estimated using the mean pairwise coefficient of kinship (*f*), representing the probability that two genes taken at random from different individuals are IBD (Falconer and Mackay, 1996). A high average kinship in a population thus implies many identical alleles and low genetic variation. The pedigree of each animal used in the study was traced back as far as known in the herd book, with an average of eight generations. Kinships were estimated using the procedures of Meuwissen and Luo (1992). Average mean kinships (*f*) were calculated excluding self-kinship, for each group and for the two groups together. Additionally, inbreeding coefficients (*F*) were calculated, representing the probability that two genes taken at random from the same individual are IBD (Falconer and Mackay, 1996). Inbreeding of an individual was given by:

$$F_i = f_{s,d}$$

where s denotes sire and d dam. The average inbreeding coefficient was calculated for each group and for the two groups together.

SNP-based diversity

The overall genetic diversity based on SNP marker data was evaluated for group EBV_{high} and EBV_{low} using the expected heterozygosity (H_{exp}) (Falconer and Mackay,

1996), which is the most widely used parameter to measure within population genetic diversity. H_{exp} is related to kinship, while observed heterozygosity (H_{obs}) (the actual number of heterozygous animals) is more related to inbreeding (Toro et al., 2009). We initially also estimated the overall H_{obs}, but because results were very similar to H_{exp} we only presented H_{exp}.

H_{exp} was calculated for each SNP marker over the 30 chromosomes and subsequently averaged over all 30 chromosomes. We also calculated the average H_{exp} over chromosome 1-29, because we expected a difference in H_{exp} for the sex chromosome because of the different effective size of sex-linked genes (Caballero, 1995). Results were almost similar, but because of this known effect we used the overall H_{exp} over chromosomes 1-29. H_{exp} was based on the allele frequencies of the SNPs within each group. Allele frequencies were estimated by counting the number of alleles 1 and dividing them by the total number of alleles. H_{exp} per SNP within groups was calculated as:

$$H_{exp,i,x} = 2p_{i,x}q_{i,x}$$

where H_{exp} is the expected heterozygosity for marker i in group x, and p and q are the allele frequencies for marker i in group x. H_{exp} was compared for the two groups using confidence intervals (see section: test statistics for SNP-based diversity), where the difference in the H_{exp} estimates between the two groups was taken as the measure for the population difference. For both groups, we also compared H_{exp} to H_{obs}.

To indicate the differentiation between the two groups, we used H_{exp} between the two groups. In that way, we investigated to what extend the two groups genetically differ from each other. H_{exp} per SNP between the two groups was calculated as:

$$H_{exp,i} = p_{1,i}q_{2,i} + p_{2,i}q_{1,i}$$

where $p_{1,i}$ and $q_{1,i}$ are the allele frequencies for group EBV_{high} and $p_{2,i}$ and $q_{2,i}$ for group EBV_{low}. H_{exp} was averaged over all SNPs (genome-wide) and at the chromosomal level.

H_{exp} is equivalent to genomic kinships that are estimated from the between individual similarity (Hayes and Goddard, 2008). The similarity for a SNP locus between individuals is defined as 1 for homozygotes with equal alleles, 0 for homozygotes with unequal alleles and 0.5 in all other cases (Eding and Meuwissen, 2001). If we denote the frequency of homozygotes 11 in the population as z_{11} and

of homozygotes 22 as $z_{22,}$ the frequency of pairwise similarities being 1 is z_{11}^2 + z_{22}^2. If the frequency of heterozygotes is denoted as $z_{12,}$ the frequency of pairwise similarities being 0.5 is z_{12}^2 + $2z_{11}z_{12}$ + 2 $z_{22}z_{12}$. Consequently, the average similarity of the total population is equal to z_{11}^2 + z_{22}^2 + $0.5z_{12}^2$ + $z_{11}z_{12}$ + $z_{22}z_{12}$. The frequency p of allele 1 in the population is given by z_{11} + $0.5z_{12}$ and the frequency q of allele 2 by z_{22} + $0.5z_{12}$. Consequently, $1\text{-}H_{exp} = p^2+q^2 = (z_{11}+0.5z_{12})^2$ + $(z_{22} + 0.5z_{12})^2 = z_{11}^2 + z_{22}^2 + 0.5z_{12}^2 + z_{11}z_{12} + z_{22}z_{12}$ which is exactly equal to the average genomic similarity in the population. The average genomic similarity is usually transformed to genomic relationships by scaling them to the range 0-1 (Hayes and Goddard, 2008).

As genomic kinships and H_{exp} are equivalent, we only analyzed H_{exp}. Additionally, the average MAF and the percentage of fixed alleles were calculated for the whole genome and each chromosome.

SNP-based diversity within chromosome regions

Neighboring SNPs showed substantial differences in diversity, and therefore, it was difficult to recognize specific regions with higher or lower diversity based on individual SNPs. To identify differences in diversity at specific chromosome regions, H_{exp} was estimated over sliding windows with a window size of approximately 5 Mb, to smoothen the heterozygosity values. This was based on the method used by Weir et al. (Weir et al., 2005). For each chromosome, the first sliding window was identified by taking the SNPs at the first 5 Mb of the chromosome. Subsequently, the window slides over the chromosome by moving the window one SNP to the right, until the end of the chromosome was reached, maintaining the same number of SNPs in each sliding window for that specific chromosome, which was on average 92 SNPs. In that way, window size will not always be exactly 5 Mb. For each sliding window, H_{exp} was estimated by taking the average of all H_{exp} values of the SNPs lying in that sliding window.

Test statistics for SNP-based diversity

To test whether the estimated SNP-based diversity (H_{exp}) differed significantly between chromosomes, chromosome regions and the two groups, bootstrapping was used (Efron and Tibshirani, 1993). Bootstrap samples were created by repeated random sampling of loci, with replacement in the following way. In each iteration, a vector of size 47,213 (= the number of loci) was created. Each locus was associated with a number between 1 and 47,213 (index number), and the vector was filled by randomly drawing index numbers with replacement. Alleles of all 90 animals were repeated according to the frequency of the index number of their locus in the

random vector (i.e. excluded if the index number was never drawn). This resulted in a new dataset (= bootstrap sample) in which the same loci appeared more than once, while other loci were not included at all. Bootstrapping was carried out in 10,000 iterations and 95% confidence intervals were estimated by taking the 250th and 9750th value after ranking. In each iteration, the MAF, % fixed alleles and H_{exp} for both groups was calculated as well as the difference between the two groups, for the whole genome, chromosomes and chromosomal regions. In this way, the variability in the estimates within and between groups of individuals caused by differences in more or less diverse loci was quantified.

The number of iterations was determined with trial runs on chromosome 1. Between 1000 and 2000 iterations, approximately 1% of the chromosome regions changed from significantly different between the two groups to not significantly different, or the other way around. Between the 9000th and 10,000th iterations, <0.1% changed significance, and no changes were observed between 10,000 and 100,000 iterations. The mean of the 10,000 bootstrap samples was in each case almost identical to the original estimate. The distribution was symmetrical with the upper and lower confidence intervals at equal distances of the mean. Therefore, all bootstrappings were based on 10,000 iterations.

Non-overlapping, 95% confidence intervals, is a too conservative test for a significant difference, i.e. the 5% level is too high. In the bootstrapping procedure, this can be seen by comparing iterations at the 95% boundaries. If in an iteration a low value was found for diversity in group EBV_{high}, the diversity in group EBV_{low} was relatively low as well, i.e. when loci are sampled that give a relatively low diversity in one group they should not be compared in the other group with another sample of loci that give a relatively high diversity. Therefore, for a direct test, it is better to calculate the difference between the groups in each bootstrap iteration and determine confidence intervals for the difference. If this interval contains 0, the two groups can be considered not significantly different. Consequently, groups with the same loci (e.g. group 1 and group 2 at the same chromosome) were considered significantly different when the 95% confidence intervals of the difference between the two groups did not include 0. Groups with different loci (e.g. different chromosomes) were considered significantly different if their 95% confidence intervals of the means did not overlap, but one should bear in mind that this is a conservative test. 44,498 chromosome regions were tested for a significant difference between the two groups. Consequently, because of multiple testing, the 5% significant level is too high. Therefore, we use the test here only as a threshold to separate regions with a high difference in diversity between the two groups

from regions with a small difference, or an opposite difference and not as a test to identify single regions with a significant difference.

3.3 Results

Genome-wide diversity

Both pedigree- and marker-estimated diversity indicated substantial differences in diversity between the two groups. Mean pedigree kinship and inbreeding coefficient over all 90 animals was 0.089 (Table 3.2) and 0.048 (Table 3.1), respectively, with a higher pedigree kinship and inbreeding coefficient for group EBV_{high} (f=0.124 and F=0.057 for group EBV_{high}; f=0.072 and F=0.036 for group EBV_{low}; Table 3.1 and 3.2). With marker-based diversity, also a lower diversity was found in group EBV_{high}. Mean expected heterozygosity (H_{exp}) over all 90 animals was 0.311, with a higher H_{exp} for group EBV_{low} (H_{exp}=0.303 for group EBV_{high} and H_{exp}=0.312 for group EBV_{low}; Table 3.2). Confidence intervals for H_{exp} within the two groups did not overlap (0.301 to 0.304 for group EBV_{high} and 0.311 to 0.314 for group EBV_{low}), so the difference in diversity between the two groups was substantial. Based on average MAF and percentage fixed alleles, similar differences between the two groups were found. In group EBV_{low}, average MAF was higher (MAF=0.228 for group EBV_{high}, MAF=0.236 for group EBV_{low}), and percentage fixed alleles was lower (% fixed alleles=9.7 for group EBV_{high}, % fixed alleles=8.7 for group EBV_{low}).

Diversity between the groups was higher than the diversity within one or both groups (f between groups = 0.073, H_{exp} between groups = 0.316, Table 3.2), indicating that SNP-based diversity will be higher if the groups are mixed than within the separate groups.

Table 3.2. Pedigree based genetic diversity (kinship) and SNP based genetic diversity (expected heterozygosity (H_{exp}) and observed heterozygosity (H_{obs}) in two Holstein groups (high and low EBV), for chromosomes 1 to 29 (excluding the X chromosome), with diversity estimated over all animals, within each group, and between the two groups, with 95% confidence intervals based on bootstrapping for the SNP based diversity.

	Pedigree based	SNP based	
	Kinship	H_{exp}	H_{obs}
	Mean	Mean (range)	Mean
Over all 90 animals	0.089	0.311 (0.309-0312)	0.310
Within group EBV_{high}	0.124	0.303 (0.301-0.304)	0.306
Within group EBV_{low}	0.072	0.312 (0.311-0.314)	0.316
Between group EBV_{high} and EBV_{low}	0.073	0.316 (0.314-0.317)	-

Diversity at chromosome level

With marker-based diversity, we could observe the diversity in more detail across the genome. Average H_{exp} within all 90 animals varied over chromosomes from 0.298 (chromosome 24) to 0.322 (chromosome 29) (Figure 3.1).

For each chromosome, H_{exp} was higher for group EBV_{low}, but differences varied over chromosomes and were significant for 26 of the 30 chromosomes (Figure 3.2).

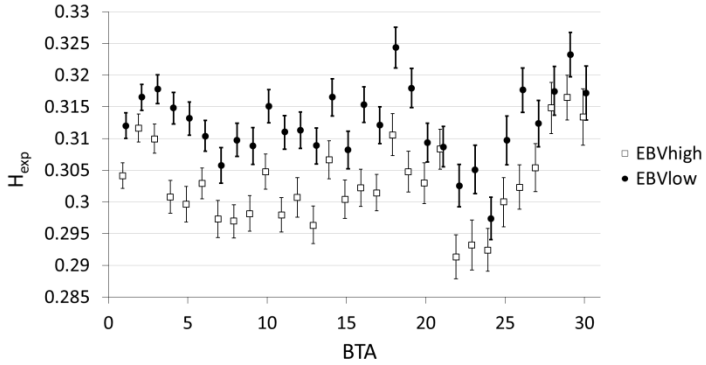

Figure 3.1. Expected heterozygosity (H_{exp}) based on SNP data in two Holstein groups (high and low EBV) over all 30 chromosomes, including 95% confidence intervals based on bootstrapping for each group for each chromosome.

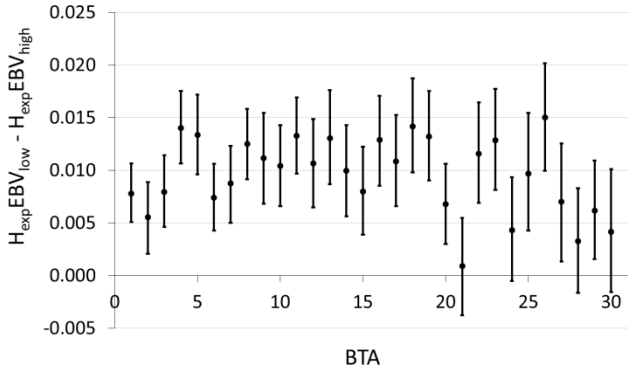

Figure 3.2. Difference in expected heterozygosity (H_{exp}) based on SNP data between two Holstein groups (high and low EBV) over all 30 chromosomes, including 95% confidence intervals based on bootstrapping for each chromosome.

Based on MAF and percentage fixed alleles, also a higher diversity was found in group EBV_{low} for most chromosomes. Largest difference in both H_{exp} and MAF, with a higher diversity for group EBV_{low}, was found at chromosome 26 (difference H_{exp}=0.0154, difference MAF=0.0148). Based on percentage fixed alleles, largest difference with higher diversity for group EBV_{low} was found at chromosome 17. For some chromosomes, diversity was higher for group EBV_{high}, but differences were very small and not always found by all three diversity measures. With percentage fixed alleles, we saw a higher diversity for group EBV_{high} at chromosome 24 (% fixed alleles=9.7% for group EBV_{high}, % fixed alleles=10.2% for group EBV_{low}), but this difference at chromosome 24 was not found with H_{exp} or MAF. Based on H_{exp} and MAF, a higher diversity for group EBV_{high} was not found. Smallest difference in H_{exp} was found at chromosome 21 (0.0003), smallest difference in MAF was found at chromosome 30 (0.0004). Thus, with SNP-based diversity, we were able to identify significant differences in diversity between chromosomes and between the two groups.

Diversity within chromosome regions

When variation was smoothed by using a sliding window of 5 Mb, differences in diversity within chromosomes and between the two groups were more clear and specific chromosome regions with significant differences could be identified.

For all chromosomes, regions with an increase or decrease in SNP-based diversity were found, with H_{exp} ranging from 0.207 (chromosome 1) to 0.393 (chromosome 30). Over the whole genome, for most of the SNPs, H_{exp} was higher in group EBV_{low}, where the difference was significant for 25.9% of the SNPs (Figure 3.3). For only 0.3% of the SNPs, H_{exp} was significantly higher in group EBV_{high}, leaving 73.9% of the SNPs with no significant difference.

The differences between the two groups varied over chromosomes, the percentage significant SNPs with a higher H_{exp} in group EBV_{low} ranged from 2.4% for chromosome 21 to 54.8% for chromosome 18 (Figure 3.3). The percentage significant SNPs with a higher H_{exp} in group EBV_{high} was much lower and ranged from 0% for 24 of the 30 chromosomes to 5.7% for the X chromosome (Figure 3.3). In comparison to the estimated diversity per chromosome, differences in diversity between the two groups were larger when diversity was estimated for specific chromosome regions, as was expected according to the lower size of SNP samples.

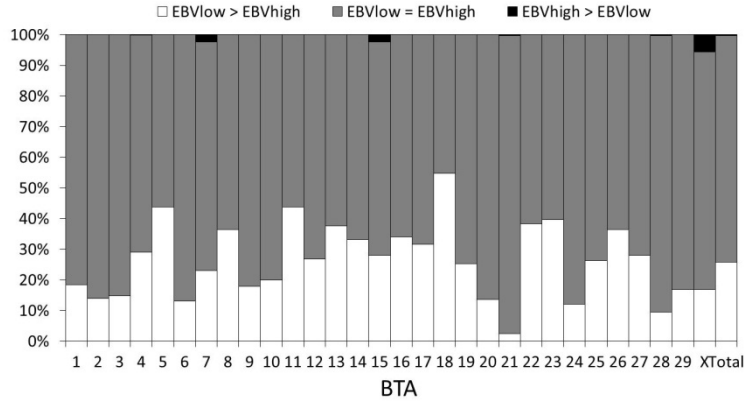

Figure 3.3. Percentage SNPs with expected heterozygosity significantly higher or lower in one of the groups, given for each of the 30 chromosomes, where significance was defined with 95% confidence intervals based on bootstrapping (not corrected for multiple testing).

The variation in diversity over chromosome regions varied substantially over chromosomes and also the differences in diversity between the two groups (Figure 3.4). Largest differences were found at chromosome 4, 7, 8, 10, 19, 26, and 30, ranging from 0.036 for chromosome 10 to 0.045 for chromosome 30 (Table 3.3, Figure 3.4). For all these ten differences, diversity was higher for group EBV_{low}. A remarkable result was the large difference at chromosome 30 that was found in the end of the chromosome, without any differences found at the rest of chromosome 30.

Table 3.3 10 chromosome (BTA) regions with the largest differences in expected heterozygosity (H_{exp}) between the two Holstein groups (high and low EBV).

BTA	Size BTA (Mb)	Peak position of the SNP (Mb)	Difference H_{exp} ($EBV_{low} - EBV_{high}$)
30	88.5	82.0	0.045
4	124.1	101.0	0.044
19	65.2	47.5	0.043
26	51.7	38.1	0.043
19	65.2	60.1	0.042
4	124.1	115.6	0.038
8	116.9	33.4	0.037
7	112.1	61.9	0.036
4	124.1	106.9	0.036
10	106.2	60.5	0.036

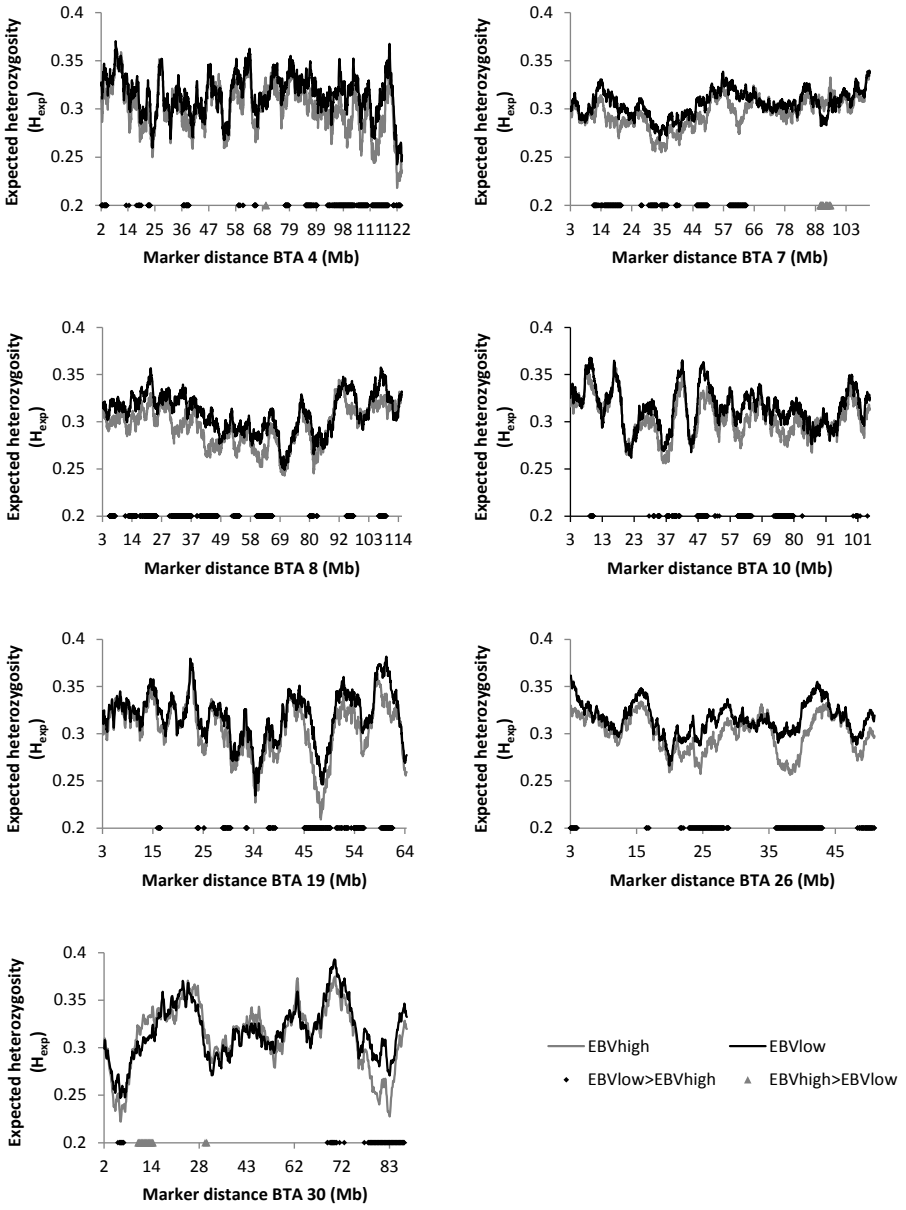

Figure 3.4. Expected heterozygosity (H_{exp}) based on SNP data for two Holstein groups (high and low EBV) for 5 Mb sliding windows at chromosomes with the largest differences between groups, with significant differences marked on the x-axis with 95% confidence intervals based on bootstrapping.

3.4 Discussion

The objective in this study was to compare pedigree- and SNP-based diversity evaluated for conservation purposes, for the whole genome, per chromosome, and for specific chromosome regions. For that purpose, two groups of Holstein animals were used that were expected to differ only slightly in diversity, i.e. represent groups of animals in a typical population that should be considered for inclusion in a gene bank. Although the two groups were genetically close, differences in diversity could be detected, with higher diversity in one of the groups. Pedigree- and SNP-based diversity gave a similar picture of diversity, but SNP-based diversity shows on which chromosome regions differences in diversity are concentrated. Both pedigree- and SNP-based diversity have been used in studies to evaluate the genetic diversity in different cattle populations all over the world (Zenger et al., 2007; Kim and Kirkpatrick, 2009; MacEachern et al., 2009; Mrode et al., 2009; Flury et al., 2010), but the current study shows that SNP-based diversity provides a more detailed picture.

In conservation of endangered breeds, populations are generally small and animals highly related and consequently differences in diversity are hard to establish. This research shows that even in these situations one can, with the help of dense marker maps, get a detailed picture of where on the genome more or less differences in diversity between populations can be found. It is, however, hard to draw conclusions on the cause of these differences nor to extend the conclusions to differences between high and low genetic merit animals in general. The differences between the two populations will have been caused by a mixture of selection, genetic drift (choice of small samples with different allelic frequencies by chance) and differences in relatedness of parents. We can, however, determine approximately the expectation of the differences in diversity given their known history.

Because the breeding values between both groups differ by around 195 EURO Inet, which is approximately ten years of selection, we assume that group EBV_{low} reflects the average Holstein Friesian population approximately ten years ago (1993), while group EBV_{high} reflects the current Holstein Friesian population (2003). With a generation interval of approximately five years, this corresponds to approximately two generations. The change in expected heterozygosity (H_{exp}) based on drift can be obtained as follows. With an N_e of 50, approximately the N_e in Holstein Friesian populations (Sørensen et al., 2005; Koenig and Simianer, 2006) and an initial $H_{exp} = 0.310$, H_{exp} two generations later equals $H_{exp,t} = H_{exp,0}(1-\Delta F)^t$ = 0.304,

where t=2 and $\Delta F = 1/(2\times 50) = 1\%$. So if we assume that both populations differ approximately two generations, we expect a difference in H_{exp} of 0.006.

One can also argue that both groups descend from the same base population with a certain H_{exp}. Theoretically, the average H_{exp} under Hardy-Weinberg equilibrium is directly related to the pedigree-based inbreeding coefficient: 1-F = H_t/H_0, where F is the pedigree-based inbreeding coefficient, H_t the observed heterozygosity in the current generation and H_0 the observed heterozygosity in the founders. As the F for, respectively, the high and low genetic merit animals was 0.06 and 0.04, this leads to H_{high}=0.94 H_0 and H_{low}=0.96 H_0, and H_{high}=0.979 H_{low}. If we calculate H_{exp} in the founder generation back from the average heterozygosity of both groups, we obtain $H_{founder}$=0.325 and H_{high}=0.306 and H_{low}=0.312. So with both lines of reasoning, the difference in H_{exp} between the two groups is expected to be small (0.006), somewhat smaller than the actual difference found (0.009). This indicates that although the difference between the two groups seems substantial using pedigree information (f=0.072 versus 0.124), and small using genomic information (H_{exp} 0.303 versus 0.312), this is actually because f and H_{exp} vary on a different scale.

Besides drift, selection may have had an effect on the differences in our study. The expected difference in H_{exp} caused by the effect of selection for Inet can be approximated as follows. Results from genome-wide association studies suggest that milk-, protein-, and fat yield are determined by many genes of small effect. Pryce et al. (2010) found 213-292 associated SNPs for each of the three traits underlying Inet and assuming approximately 100 SNPs were associated with more than one trait this leads to a total of approximately 400 SNPs associated with Inet. When assuming 400 loci of approximately equal effect, and using $\sigma_{A_{INET}}$ = 99 Euro (https://www.cr-delta.nl/nl/fokwaarden/pdf/E9.pdf) and H_{exp}=0.310, the average effect of a single locus follows from $\sigma^2_{A_{INET}} = 400 H_{exp} \alpha^2$, giving $\alpha \approx 9$ Euro. The required average change in allele frequency ($\overline{\Delta p}$) to obtain a difference of 195 Euro follows from 195 = $400\times 2\alpha \times \overline{\Delta p}$, giving $\overline{\Delta p} \approx 0.027$. The original H_{exp}=0.310 corresponds to an average MAF of approximately 0.192. When assuming that the frequent allele is favorable, i.e. that selection in the past has pulled most favorable alleles already to the higher frequency, the MAF changes to 0.192-0.027=0.165. When assuming that the frequent allele is less favorable, the MAF increases to 0.219. Thus, when 400 loci determine Inet, a change of 195 EURO should give a decrease in H_{exp} from 0.31 to approximately 0.28, or an increase to approximately

0.34. These differences are comparable with the differences in H_{exp} at several chromosome regions found in our study (Table 3.3).

In comparable studies of Sonstegard et al. (2008) and Banos and Coffey (2010), a selected and unselected Holstein line were compared using the 50 k SNP chip, where effects of selection for production traits were found on several chromosomes and chromosome regions. Also in several other studies effects of selection in Holstein cattle were observed, by comparing the Holstein breed to other dairy cattle breeds with different breeding objectives (Prasad et al., 2008; Flori et al., 2009; MacEachern et al., 2009; The Bovine HapMap Consortium, 2009). In our study we found several regions with substantial difference in diversity between the high- and low-production group, but differences could not be directly linked to specific QTL for milk production traits. However, we did not expect to find such an effect of selection for milk production in the two groups, as the groups are highly related to each other. Results show that even in small populations that are genetically close, we were able to detect differences in diversity for specific chromosome regions. This information can be very valuable when animals need to be selected for maintaining diversity in a gene bank.

It is worthwhile noting that a large part of the X chromosome did not show differences in diversity between the two groups. This can be explained by the fact that this part of the X chromosome is not overlapping with the Y chromosome and therefore is inherited from the dam only (Schaffner, 2004; Ellegren, 2009). Because selection in cattle is mainly on males, the effect of selection on diversity will be less at this part of the chromosome. Additionally, it is known that the effective population size for sex-linked genes is different from the effective population size of autosomes (Caballero, 1995), and therefore, a difference in diversity can be expected between the X chromosome and the other chromosomes. And also diversity estimation with pedigree kinship and expected heterozygosity based on the X chromosome can be different. In contrast to the results, the largest difference in diversity between the two groups was found at the part of the X chromosome which recombines with the Y chromosome and where selection acting on males is reflected (Schaffner, 2004; Ellegren, 2009).

From a conservation standpoint of view, an important question is whether we should concentrate efforts to conserve genetic diversity on certain regions of the genome, and if so on which regions. One option would be to differentiate between areas related to diversity in traits and non-functional diversity. Another option would be to concentrate on areas of reduced diversity or areas where diversity is disappearing. A distinction then has to be made between diversity lost by long-term processes in the past and diversity currently under threat by short-term

processes. In the population analyzed here, for example, differences in diversity between chromosomes follow the same pattern for both groups and result at least from before the recent divergence of the two groups and possibly from much longer ago. The smaller differences between the two groups within chromosomal regions must have arisen recently, e.g. selection for rare alleles associated with production may have increased diversity in the EBV_{high} group at some regions. However, in this study, we cannot distinguish regions with alleles under direct selection from other regions, because of small sample sizes and low divergence of the two groups. Moreover, one needs to keep in mind that ascertainment bias in the SNP chip used may also have influenced results. There may be, for example, relatively few rare alleles not associated with production present on the SNP chip, while these represent more the diversity in one of the two groups. The purpose of this study was to evaluate whether more detailed differences in diversity can be observed using SNP chips, not to find the cause of these differences. By combining results from this study with, for example, large scale studies of diversity between breeds (Stella et al., 2010) we may take better informed decisions on what diversity to conserve.

Here, we estimated differences in diversity between groups within a breed. SNP chips may also be used to estimate diversity between breeds. However, one should keep in mind that ascertainment bias may be substantial, especially in small endangered low-production breeds not used in the development of the SNP chip. On the other hand, SNP chips may provide information on the selection history in different breeds (Lynn et al., 2005; Worley et al., 2006; Hayes et al., 2008; MacEachern et al., 2009) and admixture between breeds using patterns of linkage disequilibrium. The challenge will be to develop methods suitable for small populations of the size as analyzed in this study.

With the availability of dense marker maps, more measures of diversity have become available, up to observing genetic diversity at specific chromosome regions. The present study shows that differences in diversity can be detected for specific regions in small populations that have recently diverged. In parts, these measures differ from the conventional pedigree-based diversity and therefore may have consequences in conservation, for example, when candidates for a gene bank are selected based on the amount of genetic diversity to be conserved. Also, with dense marker maps, we can identify already conserved diversity across the genome preserved in a gene bank and add candidates with different diversity. Thus, information now available through the use of SNP chips may lead to a different selection of animals to be conserved in the gene bank than selection based on pedigree information alone.

3.5 Acknowledgements

This study was financially supported by the Ministry of Agriculture, Nature and Food (Program “Kennisbasis Dier”, code: KB-04-002-021).

4

Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information

K.A. Engelsma[1,2], R.F. Veerkamp[1], M.P.L. Calus[1] and J.J. Windig[1,3]

[1] Wageningen UR Livestock Research, Animal Breeding and Genomics Centre, P.O. Box 65, 8200 AB Lelystad, The Netherlands; [2] Wageningen University, Animal Breeding and Genomics Centre, P.O. Box 338, 6700 AH Wageningen, The Netherlands; [3] Centre for Genetic Resources, The Netherlands (CGN), P.O. Box 65, 8200 AB Lelystad, The Netherlands

Journal of Animal Breeding and Genetics (2011) 128, 473-481

Abstract

Up to now, prioritization of animals for conservation has been mainly based on pedigree information; however, genomic information may improve prioritization. In this study, we used two Holstein populations to investigate the consequences for genetic diversity when animals are prioritized with optimal contributions based on pedigree or genomic data and whether consequences are different at the chromosomal level. Selection with genomic kinships resulted in a higher conserved diversity, but differences were small. Largest differences were found when few animals were prioritized and when pedigree errors were present. We found more differences at the chromosomal level, where selection based on genomic kinships resulted in a higher conserved diversity for most chromosomes, but for some chromosomes, pedigree-based selection resulted in a higher conserved diversity. To optimize conservation strategies, genomic information can help to improve the selection of animals for conservation in those situations where pedigree information is unreliable or absent or when we want to conserve diversity at specific genome regions.

Key words: genetic variation, conservation, pedigree, SNP, kinships, gene bank

4.1 Introduction

Because of the globalization of livestock breeds, many local breeds are threatened (FAO, 2009). To maintain genetic diversity, conservation of within and between breed genetic diversity is needed. The preferred conservation strategy is to maintain breeds in their natural environment (*in situ*), where *ex situ* conservation is recommended as a complementary strategy to conserve genetic diversity for the future. As not all animals within a breed can be conserved in a gene bank (*ex situ*), prioritization of animals is needed in such a way that as much genetic diversity as possible is conserved.

Prioritization of animals for conservation purposes is in most cases based on pedigree information. Marker information, like microsatellite data, can also be used for conservation decisions (Fernandez et al., 2005; Lenstra, 2006; Oliehoek et al., 2006; Toro et al., 2009) and can help to improve the selection of animals (Kinghorn et al., 2009). A commonly used method to select animals for conservation purposes is the optimal contribution method (Meuwissen, 1997; Grundy et al., 1998; Sonesson and Meuwissen, 2000), which minimizes the average relationship between candidates selected for the gene bank, thereby maximizing the conserved genetic diversity. By selecting those animals with the smallest pedigree relatedness, we assume that the maximum amount of genetic diversity has been conserved. Until recently, the use of optimal contributions in conservation studies is based on pedigree information.

With the introduction of dense SNP marker maps, we are able to evaluate the diversity over the genome in more detail. SNP-based diversity differs from pedigree-based diversity, as variation caused by Mendelian sampling is taken into account next to pedigree differences, giving a more precise diversity estimation. Another advantage of SNP-based diversity compared with pedigree-based diversity measures is that diversity at specific genome regions can be evaluated and monitored (Engelsma et al., 2010). When conservation decisions are based only on pedigree information, we might inadvertently loose diversity at specific genome regions. Sonesson et al. (2010) showed with computer simulations that excessive inbreeding rates at specific genome regions, especially those under direct selection, are prevented when optimal contributions are based on genomic information. However, the practical consequences for conserved diversity in a gene bank when selecting animals with pedigree or genomic information and the effect on the diversity at specific chromosomes have not been evaluated yet.

The objective in this study was to investigate the consequences for genetic diversity when animals are prioritized for conservation with optimal contributions

based on pedigree or genomic data and whether there are differences in diversity at the chromosomal level. For this evaluation, we used two datasets: a small population with 90 Holstein Friesian heifers and a large population with 566 Holstein Friesian heifers.

4.2 Material and methods

Animals

In this study, we used two Holstein Friesian populations. Both populations are a cross-section of the Holstein population present at that time in the Netherlands. The first population was larger, contained more related animals, and was of an earlier date than the second population.

The first population consisted of 566 Holstein Friesian heifers, born between 1990 and 1997 throughout the Netherlands. Breed composition of the animals was 100% Holstein Friesian. More information of this population can be found in the study by Veerkamp et al. (2000). The 566 animals were sired by 97 bulls, with a maximum of 36 individuals per bull. There were 72 sire groups with an average of 7.6 animals per sire group. There were 52 full-sib groups, with an average of 2.4 and a maximum of five animals per full-sib group; 98 dams had two or more daughters represented. There were 3433 animals included in the pedigree, with an average of 6.3 generations.

The second population was much smaller and consisted of 90 Holstein Friesian heifers, which were purchased in 2003 from 61 different farms throughout the Netherlands. Breed composition of the animals was 100% Holstein Friesian or 87.5% Holstein Friesian and 12.5% Dutch Friesian-Holstein (n=6). Further details of the used population can be found in the study by Beerda et al. (2007) and by Windig et al. (2008). The animals were sired by 49 different bulls, with a maximum of six individuals per bull. There were 19 sire groups with an average of 3.1 animals per sire group. There were no full-sib groups, and no mothers with daughters present in the population. Pedigree records were provided by the herd book, and the pedigree of each animal was traced back as far as known, with 3929 animals included in the pedigree and an average of 7.8 generations.

Genotyping

DNA was extracted from the 90 and 566 animals and used to determine genotypes at 54,001 SNP loci with the Illumina Bovine SNP50 Bead Chip (Illumina Inc., San Diego, CA) array. A SNP quality check was carried out before the analysis, and for this check, we used the two populations together as one data set. In the cleaning process, we removed SNPs without known position on the genome, SNPs for which

more than 5% of the animals had a missing genotype and SNPs with extreme deviations from Hardy-Weinberg equilibrium (chi-square test $\chi^2 > 600$) (Wiggans et al., 2009). Because we were also interested in those parts of the genome with extremely low or no variation, non-segregating SNPs and SNPs with a very low minor allele frequency (MAF) were not removed from the dataset. After all editing steps, 47,213 SNPs were left and used in the analysis.

Selection methods to prioritize animals for conservation

For both populations, a group of animals was selected with the aim to maximize genetic diversity in the selected group. Selection was done with the program Gencont (Meuwissen, 2002), which was used to calculate optimal contributions based on kinships from pedigree data ($OC_{pedigree}$) or genomic data ($OC_{genomic}$). Generally, optimal contributions are used to maximize breeding values while restricting the inbreeding rate. This is carried out under two constraints: both sexes have to contribute 50% each to the next generation, and the relatedness in the next generation is fixed to the value corresponding to the required inbreeding rate. As breeding values need not to be optimized for storage in the gene bank and the populations analyzed consisted of only females, both constraints are dropped here. Consequently, the relatedness of the animals selected for storage in the gene bank was minimized. This was done by finding the optimum contribution $\mathbf{c_o}$ that minimizes **c'Ac**, which is given by:

$$\boldsymbol{c_o} = \frac{\boldsymbol{A^{-1}1}}{\boldsymbol{1'A^{-1}1}}$$

where **c** is a vector with optimal contributions for all individuals, **A** is a matrix with kinships, and **1** is a column vector of ones (Meuwissen, 1997; Sonesson and Meuwissen, 2001; Eding et al., 2002). The vector with optimal contributions is summing to 100%, with a varying number of animals contributing to the gene bank (those with a contribution larger than 0%). Because, in practice, it is hard to store varying contributions of animals in a gene bank, optimal contributions were estimated for a predetermined number of animals with equal contributions.

To investigate the effect of the predetermined number of animals on the differences in conserved diversity, we first selected sets of 5, 10, 20, 40, 70, and 80 animals from both populations and compared the average genomic kinships in the sets as a measure of conserved diversity. Based on these results, we decided to perform a detailed comparison for sets of 10 selected animals (both populations) and additionally 58 animals for the large population. The latter being the same in

percentage of selected animals (approximately 10%) as 10 animals in the small population.

Firstly, to calculate ($OC_{pedigree}$), Gencont was used with the A matrix calculated from pedigree records (Falconer and Mackay, 1996) to minimize relatedness in the selected group. Secondly, to calculate ($OC_{genomic}$), the A matrix was replaced by the G matrix, which contained the genomic kinships between the animals within each population. Genomic kinships were estimated using similarities between individuals averaged over all SNPs (Hayes and Goddard, 2008). Similarities were calculated by the similarity index (Jacquard, 1983; Lynch, 1988; Eding and Meuwissen, 2001), written as:

$$S_{xy,l} = ¼[I_{11} + I_{12} + I_{21} + I_{22}]$$

where I_{ij} is an indicator variable which is 1 when allele *i* on SNP *l* in the first animal and allele *j* on the same SNP in the second animal are identical, otherwise it is 0. Subsequently, $S_{xy,l}$ can have three possible values: 1, ½ and 0. The average genomic kinship (f_g) was estimated by multiplying $S_{xy,l}$ by two and averaging the values over all SNPs. To compare pedigree and genomic kinships, genomic kinships were transformed by setting the smallest kinship to zero using:

$$f_g = (f_g - f_{min}) / (1 - f_{min})$$

where f_{min} is the minimum kinship in the matrix (Hayes and Goddard, 2008). In this way, both pedigree and genomic kinships could vary from 0 to 2, where kinships of 1 and higher are self-kinships. Self-kinships above 1 indicate that an animal is inbred. Thirdly, to evaluate the effectiveness of both selection methods, they were compared with random selections. These were performed 100 times in each population, where for each replicate, 10 animals for both populations and 58 animals for the large population were randomly selected. In this way, we qualified our results by obtaining information about the sampling variation of the conserved diversity in a randomly selected group.

Comparison of selection methods to prioritize animals for conservation

To compare the different selection methods, both pedigree- and SNP-based diversity measures were used. Based on pedigree data, the average pedigree kinship including self-kinship was calculated for the two populations and for the selected groups. Based on SNP data, genomic kinships (similarities), MAF and

percentage fixed alleles were estimated for the two populations and for the selected groups. As genomic kinships were calculated for each SNP marker over the genome, genomic kinships were averaged over SNPs, for the whole genome and for each chromosome. MAF and percentage fixed alleles were also calculated for the whole genome and for each chromosome, to evaluate the consequences of the selection methods for the SNP-based diversity for both the whole genome and each chromosome separately.

4.3 Results

Genome wide

The number of selected animals had a substantial effect on the conserved diversity with both $OC_{pedigree}$ and $OC_{genomic}$. Loss of diversity was largest when five or 10 animals were selected (Figure 4.1). The difference in conserved diversity between $OC_{pedigree}$ and $OC_{genomic}$ was larger when animals were selected from the large population compared with those from the small population. The difference between $OC_{pedigree}$ and $OC_{genomic}$ disappeared for the small population when more than 20 animals were selected. Selection of 10 animals from the small population and 58 animals from the large population gave a similar difference between $OC_{pedigree}$ and $OC_{genomic}$; therefore, results were investigated in more detail for these selections.

Selection of 10 animals increased average kinships and thus reduced diversity compared with the whole population. In the small population, diversity evaluated by the average pedigree kinship increased from 0.100 in the original population to 0.136 with $OC_{pedigree}$ and to 0.144 with $OC_{genomic}$ (Table 4.1). On the other hand, diversity evaluated with genomic kinships increased more with $OC_{pedigree}$, from 0.124 in the original population to 0.164, than with $OC_{genomic}$ (to 0.153). Thus, kinships are not minimized if they are based on another source of information (pedigree or genomic) than used to calculate the optimal contributions. In terms of percentage fixed alleles and MAF, conserved diversity also decreased when selecting 10 animals with $OC_{pedigree}$ and $OC_{genomic}$, but the difference between the two selection methods was negligible (8.0% and 0.235 in the original population, 14.4% and 0.230 with $OC_{pedigree}$ versus 14.5% and 0.232 with $OC_{genomic}$). There was a substantial overlap of 60% between the ten animals selected by both selection methods.

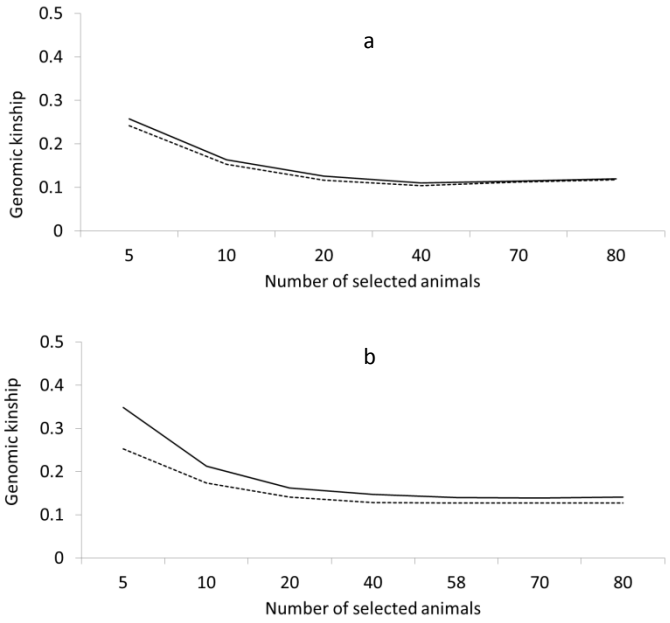

Figure 4.1 Influence of number of selected animals on the genomic kinship when selecting with optimal contributions based on pedigree kinships ($OC_{pedigree}$) (solid line) compared to genomic kinships ($OC_{genomic}$) (dotted line), for a population of 90 (a) and 566 (b) Holstein Friesian heifers and different numbers of animals being selected (5, 10, 20, 40, 58 (only for the large population), 70, and 80).

Table 4.1 Differences in genome wide diversity (pedigree kinship, genomic kinship, minor allele frequency (MAF), and percentage fixed alleles) when animals are selected with optimal contributions based on pedigree kinships ($OC_{pedigree}$), optimal contributions based on genomic kinships ($OC_{genomic}$), and random selection, for a population of 90 Holstein Friesian heifers.

	Before selection	Selection 10 animals			
		$OC_{pedigree}$ [1]	$OC_{genomic}$ [2]	Random [3]	
Diversity measures				Mean (sd)	Range
Pedigree kinship	0.100	0.136	0.144	0.184 (0.013)	0.163-0.229
Genomic kinship	0.124	0.164	0.153	0.207 (0.014)	0.182-0.251
MAF	0.235	0.230	0.232	0.224 (0.002)	0.218-0.228
% fixed alleles	8.0	14.5	14.4	16.3 (0.509)	15.2-17.7

[1] selection with optimal contributions based on pedigree kinships
[2] selection with optimal contributions based on genomic kinships
[3] random selection (100 replicates)

With random selection, average pedigree kinship in the selected group could reach 0.229, considerably higher than with optimal contribution selection. Also, genomic kinships (up to 0.251), MAFs (down to 0.218) and percentage fixed alleles (up to 17.7%) indicated that considerable losses of diversity are possible when animals are selected at random. For all four diversity measures, $OC_{pedigree}$ and $OC_{genomic}$ always did better or as good as the best run of the random sampling.

The results for the large population were comparable with results for the small population, although differences between selection methods were larger (Table 4.2). Percentage fixed alleles in the 10 animals selected was higher with $OC_{pedigree}$ (14.1%) than with $OC_{genomic}$ (13.6%). While in the small population, the two sets of animals selected with $OC_{pedigree}$ and $OC_{genomic}$ had an overlap of 60%, in the large population this overlap was only 30%.

Diversity was clearly larger when 10% of the animals in the large population (being 58 animals) were selected, instead of 10 animals (Table 4.2). The average kinship in the selection of 58 animals was in all cases even smaller than the average kinship of the whole population, which was because of the elimination of closely related animals. When comparing $OC_{pedigree}$ and $OC_{genomic}$, the differences followed the same pattern as in the selection of 10 animals. $OC_{pedigree}$ resulted in lower average pedigree kinship and higher genomic kinships, compared with $OC_{genomic}$ (Table 4.2). The difference in percentage of fixed alleles, however, disappeared. Comparing the two sets of animals selected with pedigree kinships and genomic kinships, 41% of the 58 animals in the selected groups were the same.

Table 4.2 Differences in genome wide diversity (pedigree kinship, genomic kinship, minor allele frequency (MAF), and percentage fixed alleles) when animals are selected with optimal contributions based on pedigree kinships ($OC_{pedigree}$), optimal contributions based on genomic kinships ($OC_{genomic}$), and random selection, for a population of 566 Holstein Friesian heifers.

	Before selection	Selection 10 animals				Selection 58 animals (10%)			
		$OC_{pedigree}$ [1]	$OC_{genomic}$ [2]	Random [3]		$OC_{pedigree}$ [1]	$OC_{genomic}$ [2]	Random [3]	
Diversity measures				Mean (sd)	Range			Mean (sd)	Range
Pedigree kinship	0.081	0.103	0.129	0.176 (0.013)	0.153-0.218	0.049	0.064	0.097 (0.005)	0.084-0.110
Genomic kinship	0.163	0.212	0.174	0.254 (0.014)	0.222-0.301	0.141	0.127	0.179 (0.004)	0.167-0.190
MAF	0.236	0.230	0.232	0.225 (0.002)	0.219-0.229	0.239	0.241	0.234 (0.001)	0.232-0.236
% fixed alleles	6.5	14.1	13.6	16.3 (0.579)	15.1-17.9	7.9	7.8	9.1 (0.178)	8.7-9.8

[1] selection with optimal contributions based on pedigree kinships
[2] selection with optimal contributions based on genomic kinships
[3] random selection (100 replicates)

Specific chromosomes

Selection with optimal contributions based on pedigree kinships and genomic kinships have consequences for the genetic diversity at specific chromosomes comparable to genome-wide diversity. However, both the difference between the two selection methods and the average genomic kinship varied over chromosomes (Table 4.3 and 4.4). Before selection, largest genomic kinship was found at chromosome 27 (small population) and X (large population) and smallest genomic kinship at chromosome 1 (both small and large population) (Table 4.3 and 4.4). For all selections with $OC_{pedigree}$ and $OC_{genomic}$ in both populations, highest genomic kinship was found at chromosome X and the lowest at chromosome 1. Differences between the two selection methods varied over chromosomes and also differed per population and selection. When 10 animals were selected from the small population, kinships were higher for the selection with $OC_{pedigree}$ for 23 chromosomes, with a maximum difference of 0.034 at chromosome 20 (Figure 4.2). When 10 animals were selected from the large population, differences were much larger (maximum difference of 0.174 at chromosome 22), with higher kinships for the selection with $OC_{pedigree}$ for 29 chromosomes (Figure 4.2). When selecting more animals from the large population, differences between the two methods were smaller (maximum difference of 0.025 at chromosome 10) and more consistent over chromosomes, with higher kinships for the selection with $OC_{pedigree}$ for 28 chromosomes (Figure 4.2).

Table 4.3 Differences in genetic diversity per chromosome (genomic kinship) when animals are selected with optimal contributions based on pedigree kinships ($OC_{pedigree}$) and genomic kinships ($OC_{genomic}$), for a population of 90 Holstein Friesian heifers.

	Before selection	Selection 10 animals	
		$OC_{pedigree}$ [1]	$OC_{genomic}$ [2]
Diversity measures over chromosomes	Range (sd)	Range (sd)	Range (sd)
Genomic kinship	0.195-0.367 (0.049)	0.219-0.428 (0.052)	0.195-0.414 (0.053)
Highest genomic kinship	BTA 27	BTA X	BTA X
Lowest genomic kinship	BTA 1	BTA 1	BTA 1
% fixed alleles	6.4-9.1 (0.747)	12.5-15.5 (0.824)	12.2-15.9 (0.867)

[1] selection with optimal contributions based on pedigree kinships
[2] selection with optimal contributions based on genomic kinships

Table 4.4 Differences in genetic diversity per chromosome (genomic kinship) when animals are selected with optimal contributions based on pedigree kinships ($OC_{pedigree}$) and genomic kinships ($OC_{genomic}$), for a population of 566 Holstein Friesian heifers.

	Before selection	Selection 10 animals		Selection 58 animals (10%)	
		$OC_{pedigree}$ [1]	$OC_{genomic}$ [2]	$OC_{pedigree}$ [1]	$OC_{genomic}$ [2]
Diversity measures over chromosomes	Range (sd)	Range (sd)	Range (sd)	Range (sd)	Range (sd)
Genomic kinship	0.235-0.436 (0.045)	0.401-0.589 (0.047)	0.310-0.541 (0.052)	0.195-0.405 (0.045)	0.188-0.408 (0.049)
Highest genomic kinship	BTA	BTA X	BTA X	BTA X	BTA X
Lowest genomic kinship	BTA 1	BTA 5	BTA 1	BTA 1	BTA 1
% fixed alleles	4.9-7.6 (0.655)	12.4-16.1 (1.065)	11.6-15.6 (0.944)	6.1-9.1 (0.705)	6.1-9.2 (0.761)

[1] selection with optimal contributions based on pedigree kinships

[2] selection with optimal contributions based on genomic kinships

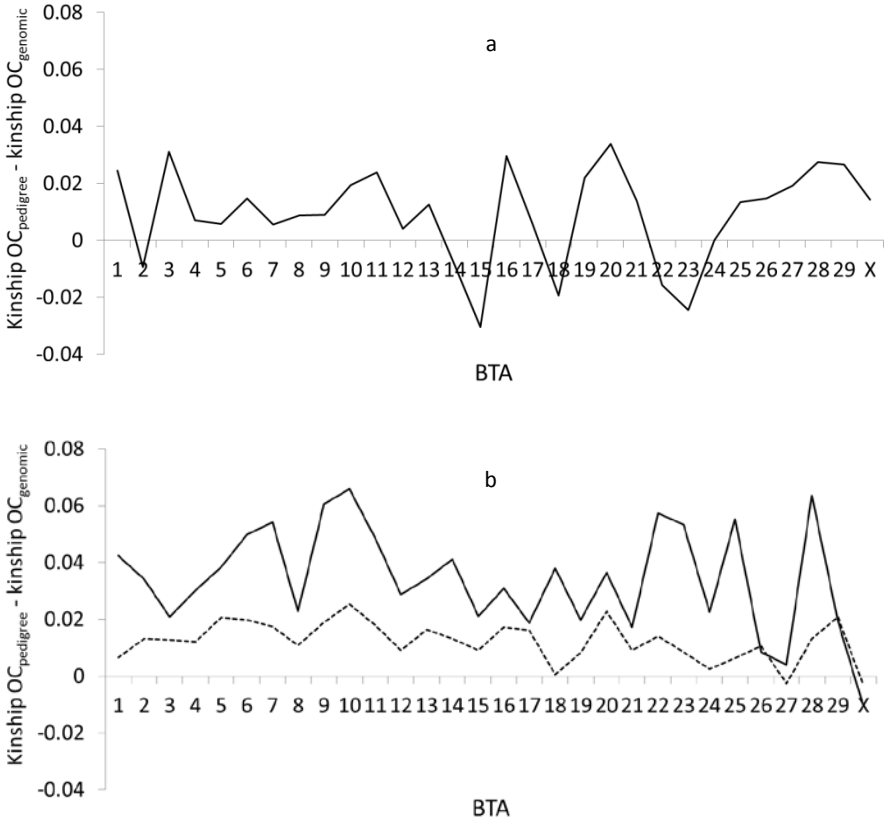

Figure 4.2 Differences in genomic kinship over chromosomes when animals are selected with optimal contributions based on pedigree kinships ($OC_{pedigree}$) or genomic kinships ($OC_{genomic}$), for (a) a selection of 10 animals from a population of 90 Holstein Friesian heifers and (b) a selection of 10 animals (solid line) and 58 animals (dotted line) from a population of 566 Holstein Friesian heifers.

The differences over chromosomes between the two selection methods were also found by observing the percentage fixed alleles, although the differences were much smaller. The percentage fixed alleles after the selection of 10 animals in both the small and the large population was twice as high compared with before selection, but differences between $OC_{pedigree}$ and $OC_{genomic}$ were somewhat larger in the large population (Figure 4.3). When selecting a larger group of animals (58 animals from the large population), the difference between $OC_{pedigree}$ and $OC_{genomic}$ was almost similar for most of the chromosomes (Figure 4.3). In contrast to the results with genomic kinship, more chromosomes with a higher percentage fixed

alleles for the selection with $OC_{pedigree}$ were found. Percentage fixed alleles was higher for selection with $OC_{pedigree}$ for 18 chromosomes in the small population, 22 chromosomes in the large population with 10 animals selected, and 20 chromosomes in the large population with 58 animals selected (Figure 4.3).

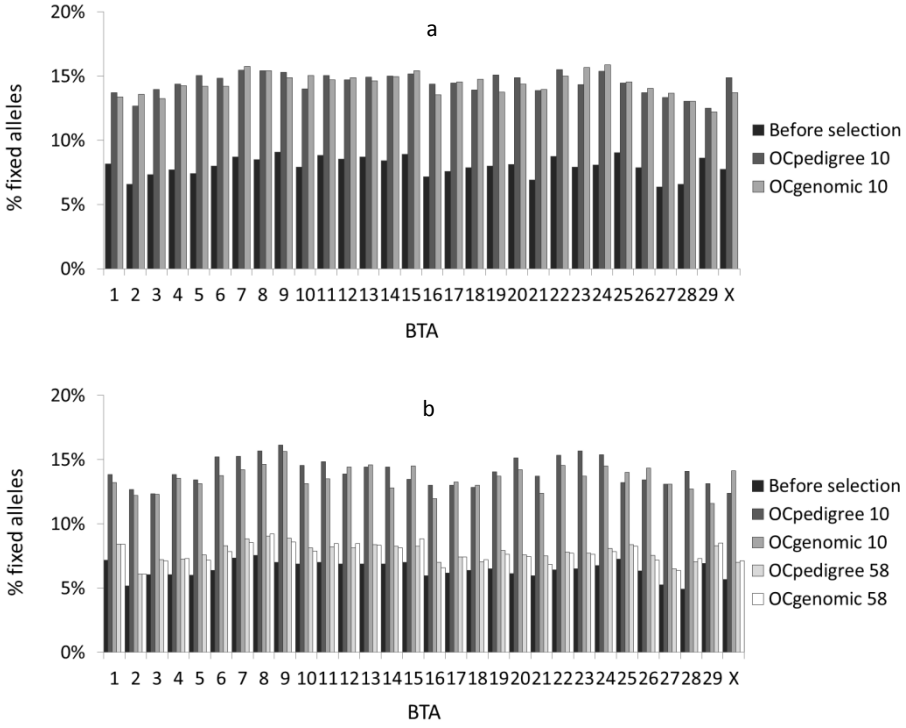

Figure 4.3 Percentage fixed alleles over chromosomes when animals are selected with optimal contributions based on pedigree kinships ($OC_{pedigree}$) or genomic kinships ($OC_{genomic}$), for (a) a selection of 10 animals from a population of 90 Holstein Friesian heifers and (b) a selection of 10 animals and 58 animals from a population of 566 Holstein Friesian heifers.

4.4 Discussion

The objective of this study was to evaluate the consequences for genetic diversity when animals are prioritized with optimal contributions based on pedigree or genomic data and to see whether there are differences between chromosomes. Differences in conserved diversity between optimal contributions based on pedigree or genomic kinships were indeed found, especially when few animals were selected from the population, but these differences varied over chromosomes. Selection based on genomic kinships resulted in a somewhat higher conserved diversity for most chromosomes, but for some chromosomes, the conserved diversity was lower, especially when animals were selected from the small population. However, when diversity was evaluated with MAF or percentage fixed alleles, differences between selections with pedigree of genomic kinships were not found. To optimize conservation strategies, genomic information can help to improve the selection of animals for conservation in those situations where a relatively small number of animals is selected, especially to conserve the diversity at specific chromosomes.

There are two causes for differences between pedigree kinship and genomic kinship: errors, in both pedigree and genotyping (e.g. swapping of DNA samples), and Mendelian sampling. Pedigree errors probably occurred in our data, especially in the large population (Figure 4.4 and 4.5). In this population, we found animals that were half sibs based on genomic kinship, but not when based on the pedigree, and the other way around. This will have had influence on the results in our study. In the small population, pedigree errors were not so obvious (Figure 4.4), and differences between selection with pedigree or genomic kinships were much smaller. Therefore, differences in the large population are partly caused by pedigree errors.

The effect of Mendelian sampling can also be a cause for the difference between pedigree and genomic kinship. Theoretically, we expect considerable variation in e.g. full-sib relationships (Hill, 1993), while using pedigree relations, we always assume that full sibs have 50% of their DNA identical by descent. SNP markers enable to estimate the actual relationship. In that way, genomic information can be used to investigate the ‘true’ diversity for each SNP and over the whole genome. As the difference between pedigree and genomic kinship was rather small for the small population, the Mendelian sampling effect seems to be smaller in our populations than because of pedigree errors.

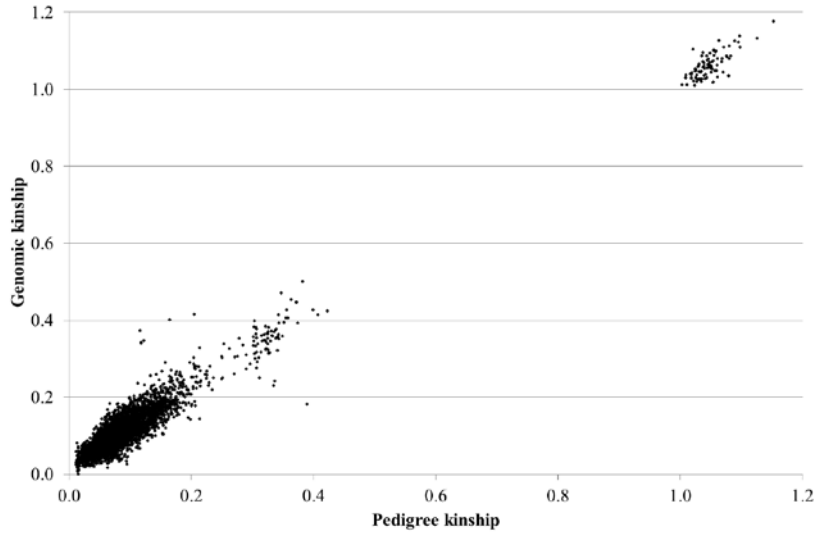

Figure 4.4 Relationship between pedigree kinship and genomic kinship for a population of 90 Holstein Friesian heifers.

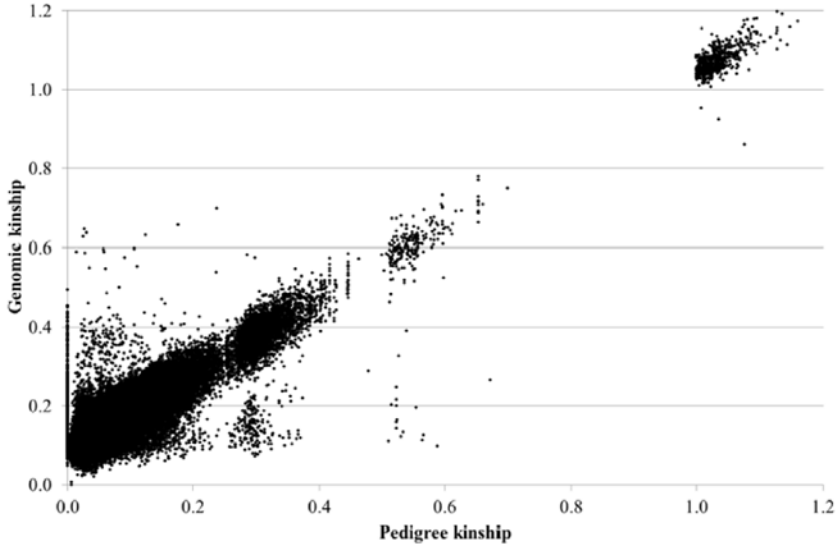

Figure 4.5 Relationship between pedigree kinship and genomic kinship for a population of 566 Holstein Friesian heifers.

The use of optimal contributions to prioritize animals for conservation purposes is expected to maintain the largest possible amount of diversity (Wray and Goddard, 1994; Weigel, 2001; Toro et al., 2009). This was also found in the current study. Selection with optimal contributions, based on either pedigree or SNP data, resulted in larger amounts of diversity to be conserved compared with random selection. In several studies, optimal contributions have been used in conservation of different breeds, for example, a sheep breed (Windig et al., 2007), a goat breed (Mucha and Windig, 2009) and a pig breed (Fabuel et al., 2004). Until now, the pedigree-based relationship matrix has been used to obtain optimal contributions. Sonesson et al. (2010) introduced the use of optimal contributions with genomic relationships based on simulated SNP data and compared this with the use of pedigree relationships to maximize genetic gain while restricting inbreeding. These results showed a higher inbreeding in regions with large Quantitative Trait Loci (QTL) underlying selected traits when using pedigree relationships instead of genomic relationships. Thus, genomic information can help in quantifying the variation in diversity and in preventing the loss of diversity at specific genome regions when using in combination with optimal contributions.

In some situations, the diversity decreased when animals were selected. For example, when 10% of the animals from the large population were selected, average genomic kinship in this selection appeared to be higher than found in the whole population. In fact, this diversity measure indicates that the animals in the selection are on average less related than in the whole population, not that the selection contains more diversity than present in the whole population. To overcome this paradox, Eding et al. (2002) developed the core set method where the diversity of a population is estimated as the average relatedness of the set of individuals with minimized relatedness selected with optimal contributions (e.g. comparable with the selection made here for a gene bank).

For practical use in a gene bank, we can make the following suggestions based on results in this paper. To select animals for genetic conservation, the use of optimal contributions is the best way to maximize the amount of diversity to be conserved. The use of pedigree information in this selection seems to be efficient, especially when animals are selected from a small population. However, in case pedigree information is unreliable or absent or when few animals are selected from a relative large population, genomic information can improve this selection resulting in more diversity within the selected group of animals. After the selection of animals with optimal contributions based on the overall diversity, genomic information can be used to identify losses of diversity at specific genome regions. To prevent this loss of diversity in the selected group, extra animals can be added

to the selection. If we would like to conserve specific genes, we might adjust the use of optimal contributions by selecting at one or more specific genes while constraining the inbreeding rate.

4.5 Acknowledgements

This study was financially supported by the Ministry of Agriculture, Nature and Food (Program “Kennisbasis Dier”, code: KB-04-002-021).

5

Consequences for diversity when animals are prioritized for conservation using the whole genome or one specific allele

K.A. Engelsma[1,2], R.F. Veerkamp[1], M.P.L. Calus[1] and J.J. Windig[1,3]

[1] Wageningen UR Livestock Research, Animal Breeding and Genomics Centre, P.O. Box 65, 8200 AB Lelystad, The Netherlands; [2] Wageningen University, Animal Breeding and Genomics Centre, P.O. Box 338, 6700 AH Wageningen, The Netherlands; [3] Centre for Genetic Resources, The Netherlands (CGN), P.O. Box 65, 8200 AB Lelystad, The Netherlands

Submitted to Journal of Animal Breeding and Genetics

Abstract

When animals are selected for one specific allele, for example for inclusion in a gene bank, this may result in the loss of diversity in other parts of the genome. The aim of this study was to quantify the risk of losing diversity across the genome when targeting a single allele for conservation when storing animals in a gene bank. From a small Holstein population, genotyped for 54,001 SNP loci, animals were prioritized for a single allele while maximizing the genome-wide diversity by using optimal contribution selection. Selection for a single allele was done for five different target frequencies, and therefore optimal contribution selection was extended with an extra constraint on the allele frequency of the target SNP marker. Results showed that elimination or fixation of alleles can result in substantial losses in genetic diversity around the targeted locus and also across the rest of the genome, depending on the allele frequency and the target frequency. It was concluded that losses of genetic diversity around the target allele are the largest when the target frequency is very different from the current allele frequency.

Key words: genetic variation, conservation, SNP, allele, kinships, gene bank

5.1 Introduction

Genetic diversity within and between breeds is under threat, and therefore conservation efforts are being made to conserve genetic diversity (FAO, 2009). To complement *in situ* conservation within the environment or production systems in which breeds were developed, *ex situ* conservation by gene banks can be used to store genetic diversity as an insurance for the future. One of the aims of a gene bank is to conserve the maximum amount of diversity possible for a given budget. To maximize genetic diversity in gene banks, prioritization of animals for incorporation in the gene bank can be done by using methods such as optimal contribution selection (Meuwissen, 2002; Fernandez et al., 2008).

In practice, conservation of animals is not only guided by optimal contributions, but often by a specific phenotype that might be affected by a single or a few genes. Examples are the conservation of the curly coat gene in horses (Thomas, 1990), the conservation of bronze turkeys (Szoke et al., 2004), and the conservation of different blood groups in cattle (Buys, 1990). With the development of molecular techniques, sometimes genotypes on a single locus can be the focus of selection as well. Examples are the elimination of genetic defects, conservation of alleles sensitive for diseases such as scrapie in sheep populations (Roughsedge et al., 2006), or fixation of alleles such as polledness in cattle (Prayaga, 2007).

When animals for a gene bank are prioritized by targeting one specific allele, this may result in the loss of diversity in other parts of the genome. This is because a relatively large part of the emphasis is put on a small part of the genome. On the other hand, due to random drift, a specific allele can be lost when animals are prioritized for conservation based on whole-genome diversity. Allele frequency may play an important role here, because alleles with low frequency have a higher chance of becoming lost compared to alleles with higher frequency. Additionally, other factors, such as population stratification, can be of influence. For example, selection for an allele that is mainly present in one family may result in a higher loss of diversity. Methods have been developed to eliminate or conserve alleles at a specific locus while controlling overall genetic diversity (Sonesson et al., 2003; Fernandez et al., 2006). An important unanswered question however is how strong the effect of allele frequency and population stratification is on the risk of losing diversity across the genome, using these methods.

The objective of this study was to quantify the risk of losing diversity across the genome when targeting a single allele for conservation when storing animals in a gene bank. In order to do so, we selected animals from a small population for inclusion in a gene bank in such a way as to maximize the genome-wide diversity

while conserving a single allele according to a target frequency. Subsequently, we investigated whether losses of diversity depended on original allele frequency of targeted alleles and stratification of the studied population.

5.2 Material and methods

To evaluate the effect of targeting single alleles when prioritizing animals for conservation in a gene bank, we used data on a Holstein cattle population which was genotyped using a 50K SNP chip. In this population we chose at random SNP markers as a target and prioritized animals accounting for both individual target alleles and overall SNP-based diversity. In order to do so the optimal contribution method was extended with an extra constraint on the allele frequency of the target SNP marker. Conserved diversity was evaluated for the region around the targeted allele and for the whole genome. Effects of population stratification and allele frequency of the targeted alleles on the loss of diversity were investigated.

Population analyzed

Data on 568 Dutch Holstein Friesian heifers that were genotyped for 54,001 SNP loci with the Illumina Bovine SNP50 Bead Chip (Illumina Inc., San Diego, CA) array was used. Characteristics of this population are given in Table 5.1, and more details can be found in Veerkamp et al. (2000). The genetic variation in this population has been analyzed in detail by Engelsma et al. (2012). SNP quality checks were done before the analysis, for which we used the current population together with another Holstein Friesian population from the Netherlands of 90 animals genotyped at the same time. In the quality check we removed SNPs without known position on the genome, SNPs for which more than 5% of the animals had a missing genotype, and SNPs with extreme deviations from Hardy Weinberg equilibrium (chi-square test χ^2>600) (Wiggans et al., 2009). After the SNP quality check, 47,213 SNPs were left for analysis. Because we wanted to investigate the effect of prioritization of a single allele with allele frequencies ranging from low to high, SNPs with a low minor allele frequency were not removed from our analysis.

Table 5.1 Characteristics analyzed population.

Number of analyzed animals	568
Breed	100% Holstein Friesian
Number of sires	97
Maximum animals per sire	36
Average number of generations	6.3
Departure from HWE	<0.001

Selection of SNPs for conservation

Our hypothesis was that allele frequency of the targeted SNP partly determines the risk of loss of diversity, because conservation of a rare allele (low minor allele frequency) is expected to result in larger changes in diversity than conservation of a common allele. To investigate the effects of allele frequency on the prioritization of animals, several SNPs with different allele frequencies were selected. We first randomly selected a position on the genome, and starting at this position we selected the first four SNPs towards the distal end of the chromosome with minor allele frequency 0.05, 0.10, 0.25 and 0.5, with a margin of +/- 0.01. In this way, selected SNPs have different frequencies with an almost similar position on the genome, so that differences in linkage disequilibrium with the rest of the genome are not due to differences in position on the genome. Random selection of positions was performed 100 times, thus in total 400 different SNPs were selected.

Target frequency

From the used Holstein Friesian population, 20 animals were selected for inclusion in a gene bank, as without constraints on a specific allele 20 animals can be selected without large losses in diversity (Engelsma et al., 2012). To investigate the effect of the target frequency for the selected allele on the genome-wide diversity, we used five target frequencies in the selection of 20 animals for the gene bank: 1) no target frequency; 2) target frequency=0.50; 3) target frequency=1 (fixation); 4) target frequency=0 (elimination) and 5) target frequency=original frequency in population. In the first scenario, 20 animals were selected without targeting the specific allele, so only the overall genetic diversity was maximized. In scenario two, the target frequency was 0.50, which meant that both alleles at the SNP should be equally represented in the selected group, and the diversity for the SNP is maximized. In scenario three and four, all selected animals should be homozygous, in scenario three for the major allele so that the minor allele is eliminated and in scenario four homozygous for the minor allele so that the major allele is eliminated. In scenario five, 20 animals were selected in order to maintain the original allele frequency found in the population.

Genomic kinships

Optimal contribution was developed to maximize breeding values while constraining the inbreeding rate to a fixed value, but can also be used to minimize the average relatedness in the next generation. The latter can be used to maximize genetic diversity of animals to be stored in a gene bank (Sonesson and Meuwissen, 2001). As we wanted to maximize the genome-wide diversity, we used kinships

estimated from genomic data. Genomic kinships were estimated by using similarities between animals, averaged over all SNPs (Hayes and Goddard, 2008). Similarities were calculated using the similarity index (Jacquard, 1983; Lynch, 1988; Eding and Meuwissen, 2001), written as:

$$S_{xy,l} = ¼[I_{11} + I_{12} + I_{21} + I_{22}] \quad (1)$$

where I_{ij} is an indicator variable which is 1 when allele *i* on SNP *l* in the first animal and allele *j* on the same SNP in the second animal are identical, otherwise it is 0. Subsequently, $S_{xy,l}$ can have three possible values: 1, ½ and 0. By averaging $S_{xy,l}$ over all SNPs, we estimated the average genomic kinship. Genomic kinships were transformed by putting the smallest kinship to zero (Hayes and Goddard, 2008).

Prioritization of animals

We used optimal contribution methods to maximize the genome-wide diversity of animals conserved in a gene bank with a pre-determined target frequency of a specific allele. With the optimal contribution method, the combination of animals with the lowest average relatedness is determined while taking one or more constraints into account. The average relatedness of animals in the gene bank is given by:

$$r_{\text{gene bank}} = \mathbf{c}\,'\mathbf{A}_{\text{population}}\mathbf{c} \quad (2)$$

where $\mathbf{A}_{\text{population}}$ is the numerator relatedness matrix, and **c** is a vector with the contributions of all animals in the population to the gene bank. The vector **c** has to sum to 1, and cannot contain negative contributions. A further constraint was that the animals in the gene bank must have the target frequency of the targeted allele, given by:

$$\mathbf{s} = \mathbf{Q}_t\mathbf{c}_t \quad (3)$$

where **s** is a vector of length 2 with the target frequencies, e.g. [1 0] if allele 1 has to be fixed in the gene bank or [0.5 0.5] if both alleles have to be equally represented in the gene bank. **Q** is a two-column matrix with the number of rows equal to the number of animals in the population. Each row contains the frequency of both alleles for each animal, i.e. [1 0] for a homozygote 1, [0.5 0.5] for a heterozygote, and [0 1] for the other homozygote 2. Originally constraint (3) was

included in optimal contribution selection to ensure that both sexes contribute equally to the next generation (i.e., **Q** was the incidence matrix of the sexes and **s** was [0.5 0.5]), but following Meuwissen and Sonnesson (2004) it can be used to reach target frequencies of a specific locus as well. The solution that minimizes $r_{genebank}$ given constraint (3) is derived by Meuwissen (1997).

To avoid negative contributions, first the optimal contributions are calculated allowing negative contributions. Next, the contribution of candidates with negative contributions are set to zero, and optimal contributions are calculated for the remaining candidates. This is repeated until all contributions are positive. Generally, the amount of genetic material (e.g. sperm doses, ova, embryos, somatic cells) produced by one animal is simply stored in the gene bank and consequently all animals selected for the gene bank should have the same contribution. Therefore, contributions above 1/n (n is the number of animals to be selected for the gene bank) are set to 1/n and optimal contributions are recalculated for the other candidates. This is repeated until all contributions are at or below the required contribution. Next the highest contribution of the remaining candidates below 1/n is fixed to 1/n after which optimal contributions are re-estimated for the remaining candidates, which is repeated until all candidates have a contribution of 1/n (Meuwissen, 1997).

The solution for the optimal contributions involves computing the inverse of **Q'AQ** (i.e. the inverse of the 2x2 matrix containing the average relatedness between and within animals carrying allele 1 and 2). However, if all animals are heterozygote, **Q'AQ** consists of four identical values and a unique inverse does not exist. Indeed if **s** ≠ [0.5 0.5] and only heterozygotes are available the constraint (3) cannot be met. However, if **s** = [0.5 0.5] the constraint might be met. When the procedures eliminating negative contributions and fixing contributions to a certain value were followed, it frequently happened that all remaining candidates were heterozygotes, and the additional candidates still to be selected required equal frequencies. In that case, the animals with the highest contributions in the previous round were selected and their contributions fixed to the required contribution. Although, one cannot be sure that in this way the combination of animals with the lowest average relatedness and the required frequency of the alleles are selected, the solution will always be close to the best solution. All calculations were performed with an adapted version of the program Gencont (Meuwissen, 2002).

Genetic diversity

To evaluate the consequences for the conserved diversity when prioritizing animals according to the different selection scenarios, genetic diversity was estimated for the original population and for the different selected groups. Genetic diversity was evaluated using the minor allele frequency (MAF), percentage fixed alleles (% fixed) and average expected heterozygosity (H_{exp}). H_{exp} was based on the allele frequencies of the SNPs (Falconer and Mackay, 1996), and calculated as:

$$H_{exp} = \frac{\sum 2p_i q_i}{n} \quad (4)$$

where H_{exp} in the selected group is the expected heterozygosity averaged over all SNPs, p and q are the allele frequencies for SNP i, and n is the number of SNPs. The expected heterozygosity averaged over the whole genome is comparable to the kinship, as it is proportional to 1 - average kinship. Besides the overall diversity, diversity was also averaged over all SNPs within a chromosome region on which the specific SNP was located. Because the selection of a group of animals for each original allele frequency was performed one hundred times, the average MAF, % fixed and H_{exp} over the 100 SNPs was taken.

Population stratification

Population stratification refers to differences in allele frequencies between subpopulations due to ancestral difference, and can be of influence when prioritizing animals for a gene bank. For example, when a specific allele is mainly present in one family and we want to conserve this allele, the prioritized animals will be related to each other and the chance of losing diversity will be higher. To investigate to what extent an SNP is confined to a genetically distinct group, we estimated the average kinship between animals with allele 1 and with allele 2, using:

$$\mathbf{f} = \mathbf{Q'AQ} \quad (5)$$

where **f** is a 2x2 matrix with the expected kinship within and between the animals with allele 1 and with allele 2. **Q** is a two column matrix as in Equation 3, and **A** is the numerator relationship matrix. We used the between-group kinship as a measure of the population stratification for each selected SNP.

5.3 Results

SNP selection

For each selected region, we were able to select four SNPs with allele frequency 0.05, 0.10, 0.25 and 0.50 lying close together. The in total 100 selected regions had a distance between the first and last selected SNP ranging from 0.32 to 3.45 Mb, with an average of 1.87 Mb and standard deviation of 0.83 Mb.

Achieved conservation goals

It was not always possible to select 20 animals obeying the target frequency for the SNP. For alleles with an original minor allele frequency of 0.05 and 0.10, no or only a few animals were homozygous for the minor allele, so that selection of 20 animals containing only the minor allele was impossible. The selection procedure also failed in one of the 100 cases for SNPs with an original minor allele frequency of 0.05 and a target frequency of 0.50, and in 5 cases to fix the major allele of SNPs with a minor allele frequency of 0.05. In all other cases the target frequency was met.

In those cases where 20 animals could be selected according to the target frequency, the genotype frequencies were generally close to Hardy Weinberg equilibrium. E.g. selection for SNPs with original and target minor allele frequency of 0.05 resulted in a percentage heterozygotes of 8.5%. Selection for SNPs with allele frequency 0.25/0.75 resulted in a percentage heterozygotes somewhat higher than 2*pq* (41.9% for 0.25/0.75 where 37.5% is expected, and 56.5% instead of 50% for 0.50/0.50).

Genetic diversity

In the whole population, averaged over the whole genome (47,213 loci), H_{exp} was 0.313, percentage fixed alleles was 6.1% and MAF was 0.236. When 20 animals were prioritized to maximize the overall genetic diversity without targeting a single allele, there was a slight increase in diversity compared to the whole population when estimated with H_{exp} and MAF. H_{exp} and MAF both were slightly higher (H_{exp}=0.315, MAF=0.238), while on the other hand more alleles were fixed (10.1%) (Table 5.2). When single alleles were targeted, diversity in the gene bank was generally smaller than in the whole population (Table 5.2). Largest loss in genome-wide diversity was found when targeting SNPs with a minor original allele frequency of 0.05 and a target frequency of 0.50 (H_{exp}=0.309, percentage fixed alleles=11.2% and MAF=0.234), and for a SNP with original minor allele frequency of 0.25 and a target frequency of 1.00 (H_{exp}=0.306, % fixed alleles=12.0 and MAF=0.231). In general, the larger the difference between the original and the

target frequency, the larger the loss in genome-wide diversity. When the target frequency was equal to the original frequency, the genome-wide diversity was the same or almost the same as for selection of animals without a target frequency.

Table 5.2 Consequences for genome-wide diversity (minor allele frequency (MAF), percentage fixed alleles and expected heterozygosity (H_{exp}), when animals are prioritized for specific SNPs with different allele frequencies.

		Target frequency prioritized SNP			
	No target	Fixate major allele	Original	0.50/0.50	Fixate minor allele
Original frequency SNP 0.05/0.95					
H_{exp}	0.315	0.315 (0.314-0.316)	0.315 (0.313-0.315)	0.309 (0.299-0.314)	-
% fixed alleles	10.1	10.1 (9.9-10.3)	10.2 (9.8-10.4)	11.2 (10.2-13.1)	-
MAF	0.238	0.238 (0.237-0.239)	0.238 (0.237-0.238)	0.234 (0.226-0.237)	-
% convergence	100	95	100	99	0
Original frequency SNP 0.10/0.90					
H_{exp}	0.315	0.315 (0.314-0.316)	0.315 (0.314-0.315)	0.312 (0.307-0.314)	-
% fixed alleles	10.1	10.1 (9.8-10.4)	10.2 (9.8-10.4)	10.6 (10.1-11.6)	-
MAF	0.238	0.238 (0.237-0.239)	0.238 (0.237-0.238)	0.236 (0.232-0.238)	-
% convergence	100	100	99	100	0
Original frequency SNP 0.25/0.75					
H_{exp}	0.315	0.314 (0.313-0.315)	0.315 (0.314-0.316)	0.314 (0.313-0.316)	0.306 (0.300-0.311)
% fixed alleles	10.1	10.2 (9.9-10.7)	10.2 (9.8-10.4)	10.2 (9.9-10.5)	12.0 (10.8-13.5)
MAF	0.238	0.238 (0.237-0.239)	0.238 (0.238-0.239)	0.238 (0.237-0.239)	0.231 (0.226-0.235)
% convergence	100	100	100	100	100
Original frequency SNP 0.50/0.50					
H_{exp}	0.315	0.312 (0.309-0.314)	0.315 (0.314-0.316)	0.315 (0.314-0.316)	0.312 (0.309-0.314)
% fixed alleles	10.1	10.7 (10.1-11.4)	10.1 (9.8-10.4)	10.1 (9.8-10.3)	10.8 (10.2-11.4)
MAF	0.238	0.236 (0.234-0.238)	0.238 (0.237-0.239)	0.238 (0.237-0.239)	0.236 (0.234-0.238)
% convergence	100	100	100	100	100

Genome-wide diversity in selected animals also varied between replicates with the same original and target frequency. The range of the genome-wide diversity was generally limited (H_{exp} between 0.313 and 0.315, Table 5.2), and ranges were larger when original and target frequencies differed more (e.g. H_{exp} 0.300-0.311 for 0.25 original frequency and 1.00 target frequency).

Besides the genome-wide diversity, we also evaluated genetic diversity within the chromosome region around the targeted SNP (Table 5.3). Before selection, the average estimated diversity around the targeted SNPs was larger than the genome-wide diversity (0.316 versus 0.313). Selected chromosome regions were thus on average more diverse than expected, despite the random selection of the 100 different chromosome regions. Apparently, the extension of the chromosome regions until it included loci with allele frequencies of 0.05, 0.10, 0.25 or 0.50 selected more diverse regions.

H_{exp} was 0.316 on average, % fixed alleles was 5.8% and MAF was 0.240. As for genome-wide diversity, diversity in the chromosome region was slightly higher in the selection of 20 animals when estimated with H_{exp} and MAF (H_{exp}=0.317, MAF=0.240), but lower when estimated with % fixed alleles (9.2%) (Table 5.3). Selection according to a target frequency resulted in larger gains or losses of diversity. Diversity in the chromosome region was highest when the target frequency was 0.50 (e.g. when diversity for the target allele was maximized). Loss of diversity in the chromosome region was largest when alleles were fixed or eliminated, and more so when the target frequency differed more from the original frequency. The largest loss of diversity around the targeted SNP was found when the minor allele of a SNP with a frequency of 0.25 was fixed (H_{exp}=0.300, % fixed alleles=13.1 and MAF=0.227) (Table 5.3).

For all scenarios, there was a large range in diversity across replicates. This range of the diversity around the targeted SNP after selection was much larger compared to the range of the genome-wide diversity, for all scenarios. In each case, replicates with a clear loss of diversity occurred as well as replicates with a clear gain in diversity. E.g. for an original frequency of 0.25 and a target frequency of 1.00, the range for H_{exp} was 0.248-0.345, for % fixed alleles 2.4-23.3%, and for MAF 0.181-0.266 (Table 5.3). Thus, targeting a specific SNP for conservation may result in a considerable loss or gain of diversity around the targeted allele.

Table 5.3. Consequences for diversity within one chromosome region (minor allele frequency (MAF), percentage fixed alleles and expected heterozygosity (H_{exp}), when animals are prioritized for specific SNPs with different allele frequencies.

		Target frequency prioritized SNP			
	No target	Fixate major allele	Original	0.50/0.50	Fixate minor allele
Original frequency SNP 0.05/0.95					
H_{exp}	0.317	0.316 (0.266-0.359)	0.316 (0.264-358)	0.317 (0.275-0.360)	-
% fixed alleles	9.2	10.5 (3.2-16.3)	9.4 (3.2-15.2)	10.2 (0.8-16.2)	-
MAF	0.240	0.239 (0.195-0.274)	0.239 (0.191-0.273)	0.242 (0.204-0.278)	-
Original frequency SNP 0.10/0.90					
H_{exp}	0.317	0.315 (0.265-0.359)	0.316 (0.273-0.364)	0.318 (0.276-0.361)	-
% fixed alleles	9.2	10.5 (1.6-16.5)	9.5 (1.6-16.9)	9.9 (2.4-17.2)	-
MAF	0.240	0.238 (0.191-0.275)	0.239 (0.197-0.279)	0.242 (0.201-0.278)	-
Original frequency SNP 0.25/0.75					
H_{exp}	0.317	0.313 (0.266-0.356)	0.317 (0.267-0.360)	0.319 (0.275-0.360)	0.300 (0.248-0.345)
% fixed alleles	9.2	10.7 (1.6-17.5)	9.4 (0.8-15.2)	9.3 (0.8-15.5)	13.1 (2.4-23.3)
MAF	0.240	0.237 (0.193-0.272)	0.240 (0.196-0.275)	0.243 (0.203-0.280)	0.227 (0.181-0.266)
Original frequency SNP 0.50/0.50					
H_{exp}	0.317	0.307 (0.260-0.341)	0.318 (0.267-0.359)	0.318 (0.257-0.360)	0.306 (0.258-0.351)
% fixed alleles	9.2	11.4 (4.8-16.9)	9.4 (1.6-16.3)	9.3 (1.6-16.3)	11.8 (3.2-21.3)
MAF	0.240	0.232 (0.191-0.265)	0.241 (0.193-0.274)	0.241 (0.184-0.274)	0.231 (0.182-0.263)

Population stratification

Differences in loss or gain of diversity when conserving animals according to a target frequency of an SNP can be partly explained by population stratification. Kinship between animals with and without the targeted allele varied considerably. For SNPs with a low minor allele frequency (0.05 and 0.10) variation in kinship was larger (range from 0.146 to 0.171) compared to SNPs with higher minor allele

frequency (range from 0.158 to 0.163), with lower kinship values indicating more population stratification.

The effect of population stratification was largest when conserving a SNP according to an allele frequency different from the original frequency of the SNP in the population. The correlation between loss of genome-wide diversity and kinship was 0.66 when conserving SNPs with minor allele frequency 0.05 or 0.10 and target frequency 0.5 (Table 5.4, Figure 5.1). In this case, conserving alleles with small allele frequency and a larger population stratification resulted in a smaller loss of diversity. This result can be expected, as in this situation the more unique animals will be selected. The opposite effect was found when the major allele was fixed for alleles with a MAF between 0.05 and 0.25. Here the correlation between loss of genome-wide diversity and kinship was negative (-0.50 to -0.54), i.e. conserving alleles with a larger population stratification resulted in a larger loss of diversity (Figure 5.2).

Table 5.4 Pearson correlation coefficients for the relation between population stratification (expressed as kinship between groups of animals with and without the targeted allele) and loss of genome-wide diversity (expressed as the difference in H_{exp} between before selection and after selection).

	Target frequency prioritized SNP			
Frequency of prioritized SNP	Fixate major	Original	0.50/0.50	Fixate minor
0.05/0.95	-0.54	-0.25	0.66	-
0.10/0.90	-0.50	0	0.58	-
0.25/0.75	-0.53	-0.28	0.36	0.61
0.50/0.50	0	-0.13	-0.25	-0.25

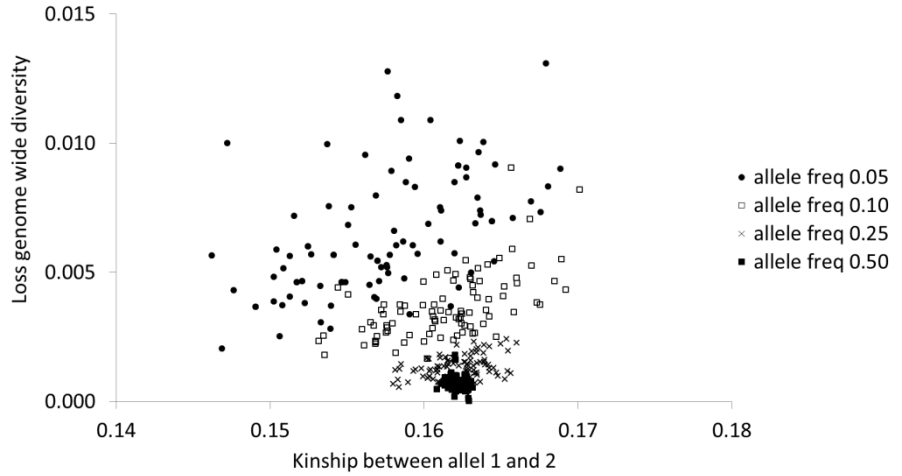

Figure 5.1 Effect of population stratification (given as the kinship between groups of animals with and without the targeted allele) on the loss of genome-wide diversity (given as the loss in H_{exp}) due to prioritizing animals for a single allele, with target frequency 0.50.

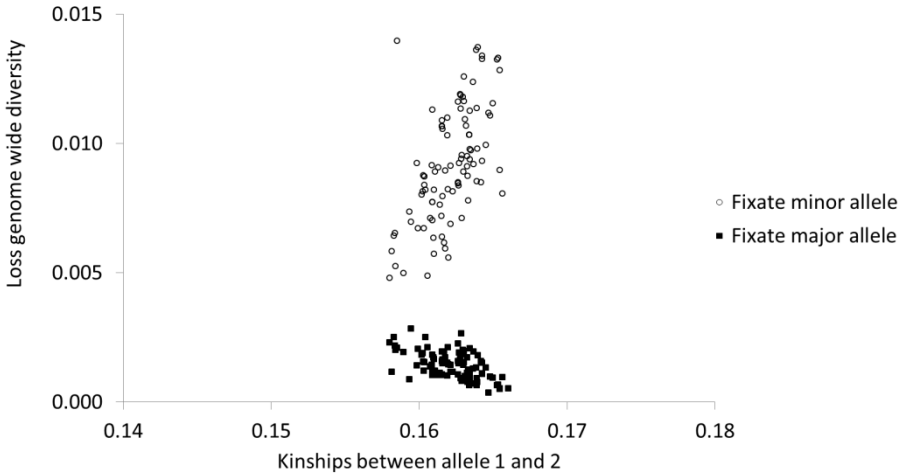

Figure 5.2 Effect of population stratification (given as the kinship between groups of animals with and without the targeted allele) on the loss of genome-wide diversity (given as the loss in H_{exp}) due to prioritizing animals for a single allele, given for SNPs with allele frequency 0.25/0.75 for which the minor or the major allele was fixated.

5.4 Discussion

It is important to maximize genetic diversity in gene banks. Targeting one specific allele when prioritizing animals for conservation can, however, result in less diversity stored in the gene bank. We showed in this study that the loss of diversity depends on the difference between the original allele frequency in the population and the target frequency in the gene bank. The largest average loss in genome-wide diversity that was found was 0.009, which can be compared to two generations of intensive selection in a Holstein population (Engelsma et al., 2012). Loss in single replicates and around the targeted allele could be much higher.

In this study, we forced the allele frequency of conserved animals to a specific value within one generation. This proved to be impossible when we tried to fix the minor allele of SNPs with minor allele frequency 0.05 or 0.10, because there were not enough homozygous animals with the targeted allele to achieve fixation. Another effect is that loss of diversity is higher than necessary, because there is less space to find animals contributing to diversity. If we would have used more generations of selection in our study, fixation of alleles with small minor allele frequency would have been possible. In general, rapid fixation of a desirable allele leads to a greater loss of diversity. Li et al. (2008) proposed methods to maximize long-term response while conserving diversity and controlling inbreeding, and showed that the rate of inbreeding is smaller when selection is done over more generations. A similar situation was found in the study of Windig et al. (2007), where sheep were selected for scrapie resistance. The rate of inbreeding was smaller when fixation of the favored allele was done over several generations of selection, instead of immediate fixation of the allele within one generation.

The effect of population stratification depended on the allele frequency of the conserved SNP and the target frequency in the selection. When minor alleles were fixed or increased to 50%, alleles with a lower kinship showed a smaller loss in diversity. Animals with a low average kinship with the rest of the population contribute more to overall diversity than animals with a high average kinship. With optimal contribution selection, these animals are generally selected to maximize genetic diversity. Thus, when animals with a low kinship are selected because they contain the targeted allele, the loss of diversity will be less compared to selection of animals with a high kinship. In contrast, when the major allele is fixed, alleles with a high kinship showed a smaller loss in diversity. This is because the minor allele is at the same time eliminated. To avoid losses of diversity when selecting animals based on a single allele, it is thus important to first determine which animals would be selected without targeting the specific alleles. If the consequence

of targeting a single allele is the loss of considerable amounts of diversity, either the target frequency should be changed, or selection should be done over more generations. E.g. if a disease allele should be eliminated, one may first mate carriers of the allele to non-carriers and select among their offspring animals without the disease allele to be included the gene bank.

Optimal contribution selection is the most powerful method to maximize genetic diversity in breeding programs (Fernandez et al., 2011), and has been used to maximize genetic gain while restricting inbreeding rate (Skaarud et al., 2011; Gourdine et al., 2012). In this study, optimal contributions were used to maximize the overall diversity and conserve a single allele according to a certain target frequency at the same time. Although losses within chromosome regions could be considerable, in general the method was successful in conserving diversity. Fernandez et al. (2006) also developed an algorithm to store alleles in the gene bank, given a certain target frequency while controlling the genetic diversity of other loci, using simulated annealing. In theory the optimal contribution method is more exact, but the advantage of the simulated annealing algorithm is that in cases where the target frequency cannot be met it can give a solution with a frequency approaching the target frequency. In practice both methods are effective in conserving genome-wide diversity while selecting for a single allele.

With optimal contribution twenty animals could be selected that were more diverse than the original 568 animals, based on H_{exp} and MAF. This is because groups of highly related animals are excluded in the selected group so that on average their relatedness is lower and H_{exp} higher. When judged by the % fixed alleles the diversity in the selected group was, however, clearly lower. The reason is that when alleles with a low frequency get fixed the average H_{exp} and MAF do not change much since the difference between a low allele frequency and zero is small.

The Holstein population that we used in this study was not a typical population normally conserved in gene banks, but the results from this study are applicable for small and/or rare populations. The Holstein population, although not under threat, has a small effective population size, comparable to breeds threatened by extinction. Therefore France, The Netherlands and the USA have started to conserve genetic material of HF bulls since the nineties. The number of HF bulls in the gene bank is 144 for France, 3755 for The Netherlands and 5013 for the USA (Danchin-Burge et al., 2011). For most livestock breeds under threat such numbers cannot be reached and hence the number of animals selected in our study (20) is more relevant for small populations to be conserved in a gene bank. The relatively small number of selected animals results in a sampling effect, which can be considerable, especially when the target frequency differs from the original

frequency in the population, as evidenced by the large variation among replicates. Moreover, we showed that population stratification may influence the loss of diversity. Especially when a small population is maintained by a few breeders in isolation from each other, population stratification may be much more pronounced than in our study. In such populations, extra care is needed when targeting a single allele in conservation.

In practice, it often occurs in conservation plans that a specific allele gets special attention, with the aim to fixate or eliminate an allele. Our study shows that this should be done with care. Results showed that elimination or fixation of alleles can result in substantial losses in genetic diversity around the targeted locus and also at the rest of the genome, depending on the allele frequency and the target frequency. Optimal contributions will limit the loss of diversity, but especially when alleles are conserved according to a target frequency different from the allele frequency loss of diversity can be considerable. Particularly in small populations it is important to be careful with selection of single alleles, because a higher LD increases the chance of losing genetic diversity around the targeted allele.

5.5 Acknowledgements

This study was financially supported by the Ministry of Economic Affairs, Agriculture and Innovation (Program “Kennisbasis Dier”, code: KB-04-002-021).

6

General discussion

6.1 Introduction

The aim of this thesis is to explore the opportunities to use SNP markers for conservation of genetic diversity within livestock breeds. We first compared two different methods to estimate genetic diversity using SNP markers (Chapter 2) and we compared the use of SNP markers for genetic diversity estimation to pedigree information (Chapter 3). Furthermore, we investigated the use of SNP markers and pedigree information for prioritization of animals for conservation of overall genetic diversity in a gene bank (Chapter 4). Finally, we investigated the use of SNP markers for conservation of a single allele (Chapter 5).

In this final chapter, the use of SNP markers for estimation and conservation of genetic diversity is discussed further, including future perspectives for conservation of genetic diversity and practical implications of the use of SNP markers for conservation *in situ* or *ex situ*.

6.2 SNP markers and conservation of genetic diversity

Technical developments in the area of molecular genetics have resulted in the availability of large numbers of SNP markers, which can now be used in genetic and characterization studies. These large numbers of SNPs give us the possibility to estimate genetic diversity in more detail, and to improve the prioritization of animals for conservation purposes. In animal breeding, the challenge is balance selection and maintenance of genetic diversity within breeding populations. In commercial breeds high selection pressure leads to a narrowing genetic base, and several small local breeds are lost every year (FAO, 2007a). Effective population sizes in animal breeds are often small, for local and endangered breeds, but also in commercial breeds like the Holstein Friesian (Goddard, 1992). The development of large numbers of SNP markers can be of help to characterize genetic diversity within breeds, and to conserve breeds in such a way that as much genetic diversity as possible is conserved. In animal breeding, SNP markers have played a prominent role in genomic selection studies, in which genomic relationships between animals are estimated. In these simulation studies generally large groups of animals were used. Therefore, the suitability of large numbers of SNP markers for genetic diversity estimation and prioritization need to be investigated in small groups of highly related animals as well.

Genetic diversity estimation with SNP markers

In Chapter 2, two methods to estimate genetic diversity were compared. The methods were based on heterozygosity of markers and on IBD probabilities. In case

of low marker densities, IBD probabilities appeared to be a better predictor for genetic diversity, because they correct for markers being alike in state (AIS) but not identical by descent (IBD). In estimating the probability of IBD for long stretches of the genome, the probability of recombination between markers was taken into account. Especially in parts of the genome without or with only a few markers this resulted a better prediction of diversity. When high density markers were used this advantage disappeared, and the two methods gave similar results. Therefore, further on in this thesis, heterozygosity was used for the estimation of genetic diversity.
With SNP markers we are able to recognize differences in genetic diversity within and between small and closely related populations. This was demonstrated for two small Holstein populations in Chapter 3. With both pedigree and SNP marker information, a difference in overall genetic diversity was found between the two populations. Estimations based on the relationship between pedigree and SNP based diversity showed that these differences were comparable to each other. Therefore, when pedigree information is not available or of bad quality, SNP markers are very suitable to estimate the genetic diversity within a population. But SNP markers can also have an advantage above pedigree information. First of all, with pedigree information we can estimate the expected relatedness between animals, but with dense SNP markers we can observe the true relatedness. For example, the pedigree relatedness between full sibs will always be 0.5, while based on SNP markers this relatedness will vary approximately between 0.4 and 0.6 (Visscher et al., 2006). The latter is more realistic, because due to sampling events full sibs can be somewhat more or less related to each other. For less close family members the variation can become larger, because every generation there is a chance that Mendelian sampling causes variation. Another advantage of SNP markers above pedigree information is that we are able to identify many more differences in genetic diversity at specific parts of the genome between the two populations, as was shown in Chapter 3. This can be useful, for example when populations have to be conserved in a gene bank. Especially in a closely related population, animals can be very similar to each other based on pedigree relatedness, and with SNP markers we can find those animals that harbor unique genetic diversity. In that way it is possible to conserve specific genes that have favorable properties.

Conservation with pedigree and SNP information

In Chapter 3 we showed that genetic diversity estimation with pedigree and SNP marker information give similar results in terms of overall genetic diversity, although SNP based diversity may be different at chromosomal or more detailed levels. An important question is whether prioritization based on pedigree or SNP marker

information results in a different selection of animals, and a different amount of conserved genetic diversity. In Chapter 4 we therefore used optimal contribution selection to prioritize animals for conservation, and we compared optimal contributions based on SNP markers to optimal contributions based on pedigree information. Optimal contribution selection is a commonly used method to prioritize animals for conservation, in combination with pedigree information (Windig et al., 2007; Mucha and Windig, 2009) or microsatellite marker data (Fabuel et al., 2004; Tapio et al., 2010). Optimal contributions in combination with SNP markers has been mainly used in selection studies for genetic improvement (Nielsen et al., 2011, Sonesson et al., 2010), but not yet in conservation studies. The results in Chapter 4 showed that with optimal contribution selection we are able to conserve more genetic diversity compared to prioritization without optimal contributions, for both pedigree information and SNP markers. The overall conserved genetic diversity was somewhat higher for prioritization with SNP data, but differences were small.

The conserved genetic diversity with SNP data increased when a small group of animals was prioritized, or when pedigree errors were present. Differences in conserved genetic diversity were larger at the chromosome level, where selection with SNP data resulted in higher genetic diversity for most chromosomes, but at some chromosomes selection with pedigree information resulted in higher genetic diversity. This means that the chance of losing diversity at specific parts of the genome is somewhat smaller when we use SNP markers. Based on these results we can conclude that optimal contribution selection based on SNP markers is a good method to improve prioritization of animals, especially when a small number of animals is selected.

Finding a balance between genome wide diversity and single allele diversity

In some situations it is desirable to conserve a specific gene that represents a favorable trait or eliminate a gene that causes a disease. Recent examples are conservation of genes responsible for polledness in cattle (Medugorac et al., 2012), or elimination of genes responsible for scrapie in sheep (Roughsedge et al., 2006), BLAD in cattle (Nagahata, 2004) and dwarfism in horses (Orr et al., 2010). However, eradication or fixation of genes incurs the risk of losing other important genes that are linked either through linkage or pleiotropic effects. Therefore, during selection of a single allele it is of importance to prevent loss of genetic diversity at the rest of the genome. The consequences of selecting one specific allele for the genetic diversity at the rest of the genome was investigated in Chapter 5. We prioritized animals for conservation in a gene bank for a single allele in combination with

maximizing the overall genetic diversity, based on SNP marker information. In order to do this, we extended the optimal contribution selection with an extra constraint on allele frequency. Results showed that elimination or fixation of alleles can result in substantial losses in genetic diversity around the targeted locus and also at the rest of the genome, depending on the allele frequency and the target frequency. Losses of genetic diversity around the target allele are the largest when the target frequency is very different from the current allele frequency. This is the so called hitchhiking effect: selection for a single allele also changes the allele frequencies of linked neighboring alleles (Smith and Haigh, 2007). These hitchhiking effects were also seen in the study of Pedersen et al. (2010), where marker assisted selection for a specific QTL resulted in extra losses of genetic diversity around the QTL. Besides the allele frequency, also the LD between the targeted allele and neighboring alleles will be of influence on the loss of genetic diversity. In case LD is strong, the chance of losing neighboring alleles will be higher. Also when the targeted allele was affected by recent mutation, the LD block around the allele will be larger, resulting in a higher chance of losing genetic diversity around the targeted allele. Therefore, especially in small populations it is important to be careful with selection of single alleles, in particular when the selected alleles have been recently arisen. However, when animals are prioritized for a single allele, it is also possible to increase the genetic diversity at other parts of the genome. When we would change the allele frequency from 0.05 to 0.5, we increase the genetic diversity for the target allele and also in the region around the allele and at the rest of the genome.

Selection strategies can result in loss of genetic diversity when we prioritize animals for alleles with a target frequency very different from the original allele frequency. Consequently, strategies are needed to reduce this loss of genetic diversity by increasing the genetic diversity around the target allele. This can be done by using optimal contribution selection with an additional constraint on genetic diversity around the target gene. This would mean that we prioritize a group of animals with the objectives to 1) conserve the single allele, 2) conserve the genetic diversity in the region around the single allele, and 3) conserve the genetic diversity over the whole genome. It is also possible to use SNP markers to prioritize animals for a trait influenced by several genes. For example, we might want to conserve genetic diversity for a trait like milk quality, and therefore we have to conserve many alleles positioned in different areas of the genome. In order to do this we have to conserve all SNP markers that are close to these milk quality genes, together with maximizing the overall genetic diversity. In theory this seems to be possible with optimal contribution selection, but it will be a challenge to run the software without mathematical problems or difficulties when we would use optimal contribution

selection with one constraint for each allele. An alternative is to use one constraint for all alleles together. However, with one constraint for all alleles it might become difficult for the program to find a solution. Another strategy is to use more than one generation to achieve the prioritization goal. By allowing more generations, the required change in allele frequency in each generation is smaller. In addition, by taking more time to achieve the prioritization goal it is possible that the variation in the population increases due to recombination.

6.3 Future perspectives for genetic diversity conservation

In the expert group reports in the Strategic Research Agenda 2011 of the Fabre Technology Platform (Fabre Technology Platform, 2011), a number of important research priorities for conservation of genetic diversity in the future have been identified. In the next section, the two most challenging research priorities have been selected and discussed: balance between genomic selection and loss of genetic diversity, and using whole genome sequence information for conservation of genetic diversity.

Balance between genomic selection and minimizing loss of genetic diversity

The intensive selection in animal breeding in the last century has made a large contribution to improvements in animal production (Fabre Technology Platform, 2011). In the last decade, new breeding methods like marker assisted selection and subsequently genomic selection have come into use to further increase genetic improvement in animal production. Especially genomic selection may be very successful in animal breeding, as nearly all genetic variance of a trait can be explained by markers covering the whole genome, resulting in genomic breeding values (Goddard, 2009). Although the intensive selection in animal breeding has resulted in successful improvement of animal production, it also caused increased inbreeding within several breeds. For example in the Holstein Friesian breed, the heavy use of a few popular AI sires has resulted in rapid genetic improvement, but also in a reduction of the effective population size (Goddard, 1992) and an increase in level of inbreeding (Thompson et al., 2000). On the other hand, it is questionable whether the genetic variation in Holstein Friesian cattle has decreased too much due to selection, since observed heritabilities of production traits have not decreased. An important question is what the effect of genomic selection is on genetic diversity within breeds, and if it is possible to apply genomic selection while restricting the loss of genetic diversity during selection.

The effect of genomic selection on genetic diversity can be both positive and negative. Genomic selection can result in higher losses of genetic diversity compared to traditional breeding methods, because animals can be selected earlier in life. Selection of younger animals results in a shorter generation interval, and subsequently in an increased rate of inbreeding per year (De Roos et al., 2011). On the other hand, genomic selection can also result in lower losses of genetic diversity compared to traditional breeding methods. With genomic selection, Mendelian sampling effects can be estimated more accurately, which enables differentiation between sibs without performance recordings and therefore reduces co-selection of sibs (Daetwyler et al., 2007; Buch et al., 2012; Pryce and Daetwyler, 2012). In that way, rather than all sibs of a family, only the best animal within a family is selected, and more families can be sampled within a breeding program, which reduces the rate of inbreeding per generation. This problem of co-selection of sibs is expected to be less relevant for dairy cattle compared to species like chickens or pigs, because family groups in dairy cattle are much smaller. One might argue that genomic selection can also increase co-selection of sibs and subsequently the rate of inbreeding, because genomic breeding values at least partly rely on recent relationships (Habier et al., 2007). To further investigate the effect of genomic selection on co-selection of sibs, correlations between genomic breeding values of sibs should be compared to the expected values. The rate of inbreeding can be positively or negatively influenced by breeding program parameters like selection intensity, number of selected animals, age of the selected animals or number of parents used. For instance, one extreme scenario could be that genomic selection is used to screen entire populations, to try and avoid selecting highly related animals by just sampling a limited number of successful families.

Application of genomic selection may especially increase inbreeding around selected alleles as a result of hitchhiking (Pedersen et al., 2010). In this thesis in Chapter 5 we also demonstrated that prioritization of animals for a single allele can result in substantial loss of genetic diversity in the regions around the selected allele. On the other hand, selection for diversity at a specific locus may result in more diversity in the region around the locus. These observations pose an important challenge for optimal applications of genomic selection: on the one hand we want to minimize the genetic diversity for genes that are beneficial for the trait (i.e. fix the favorable alleles), but on the other hand we want to maximize the genetic diversity for the rest of the genome. A method to restrict the loss of diversity during selection is optimal contribution selection, which has been developed by Meuwissen (1997) and tested in several populations (Avendaño et

al., 2003; Kearney et al., 2004; Sørensen et al., 2008). Its objective is to maximize genetic gain while restricting the rate of inbreeding by constraining the average relationship among selection candidates. Optimal contribution selection has been mainly used in traditional breeding programs based on pedigree-based relationships. Optimal contribution selection can be used as well in genomic selection breeding schemes, to reduce the loss of diversity. Sonesson et al. (2010) used optimal contribution selection based on genomic relationships, which resulted in lower overall inbreeding rates than when pedigree relationships were used. The study showed that optimal contribution selection based on pedigree relationships results in high rates of inbreeding around the selected QTL. Selection for a QTL will always result in a certain increase of inbreeding around the QTL, however, with optimal contribution selection based on genomic relationships we have more possibilities to reduce this inbreeding.

A point of debate could be whether or not the loss of genetic diversity around selected favorable alleles is important, because this loss might be very small. When genomic selection would be used for the selection of only one trait that is representing a small number of genes, loss of genetic diversity might be a problem. But in genomic selection the breeding goal often contains several traits, and together they may easily represent more than 1000 genes. In that case, after one generation of selection the change in allele frequency will probably be very small on a single locus, and therefore the loss of genetic diversity around these alleles might also be very small. However, when linkage disequilibrium is high around certain selected alleles, you might fixate haplotypes which results in higher losses of genetic diversity. Especially in small populations the LD over the genome is larger, and therefore genomic selection might lead to higher losses of genetic diversity compared to large populations.

Genomic selection can help to reduce the loss of rare alleles during selection, focusing on the long-term response to selection. According to Goddard (2009), putting extra weight (i.e. extra selection pressure) on alleles with low allele frequency is necessary in order to get an optimal long term response to genomic selection. This in contrast to maximizing the short-term response, which leads to stronger selection of alleles with large effects and with less weight on rare alleles (Bijma, 2012). Because most alleles occur at extreme frequencies, putting more weight on rare alleles will eventually result in a higher increase in the total genetic variance and long-term gain (Bijma, 2012). Goddard (2009) developed a genomic selection model in which selection of specific alleles is optimized by defining the target frequency. This was achieved by adding weights to favorable alleles, and vary these weights according to the allele frequency in the population by making

them proportional to 1/√p(1-p). Subsequently, a favorable allele with a low frequency gets a relatively high weight compared to other favorable alleles with higher frequencies, which increases the frequency of rare favorable alleles. The same genomic selection model was used in a simulation study for an inbred crop by Jannink (2010), who also showed the importance of placing additional weights to rare alleles. Li et al. (2008) used a genomic selection strategy in which also more emphasis is put on rare alleles, but they used the negative value of the logarithm of the frequency of the favorable alleles. It should be noted that in the model of Goddard (2009), however, genetic drift is not taken into account. Bijma (2012) emphasized that drift cannot be ignored because Mendelian sampling and recombination is outside the breeder's control, and therefore he suggested to put even more weight on rare alleles to reduce the probability of losing rare alleles by chance. Since the impact of drift is more important in small populations, the relevance of putting more weight on rare alleles is higher in small populations.

Goddard (2009) suggests that we also might have to put small weights on markers without any known effect, in order to prevent the loss of rare alleles at those loci. This could be done by inclusion of a polygenic effect in the model and thereby putting some selection pressure on unidentified QTL. Alternatively, a polygenic effect can be included by taking into account all markers and not only a subset. This would mean that use of genomic selection methods that use variable selection (e.g. BayesB; Meuwissen et al. (2001)) are unfavorable when the goal is to optimize long term genomic selection response, because BayesB uses only part of the SNPs to explain the variation. Methods like GBLUP would be more suitable, as this method uses all SNPs for explanation of the variation. A conclusion could be that methods that are more close to the infinitesimal model fit better the long term optimal selection response compared to finite locus methods like BayesB that fit better short term optimal selection response. In order to test this conclusion, different genomic prediction methods should be compared, looking at their effect on inbreeding and the loss of genetic variation over generations. This was, for example, done in the study of Bastiaansen et al. (2012), where GBLUP was compared to two methods that are based on a smaller number of SNP markers, a Bayesian method and partial least squares regression. In this study, GBLUP resulted in 0.6% and 0.9% less inbreeding and on average a one third smaller reduction of genetic variance, which supports the conclusion that methods like GBLUP will be more suitable to reach an optimal long term response to genomic selection and at the same time reduce the loss of genetic diversity.

In summary, several options are available to simultaneously minimize loss of genetic diversity and optimize response to genomic selection. SNP effects, used for

genomic selection, can be estimated using models with weights on loci to achieve optimal long-term selection. From the currently widely applied methods to estimate SNP effects, GBLUP appears to be the model that best fits the long-term goals. In any case, regardless of the genomic prediction model used, application of optimal contribution selection is required to reduce the loss of genetic diversity around target genes and across the rest of the genome.

Added value of increased marker density and whole genome sequence data for conservation

In the near future, high density SNP data and complete genome sequences will be available for most livestock species and breeds at affordable prices. It is the question whether this extra information can increase the level of conserved genetic diversity. For conservation of overall genetic diversity, a higher SNP density can improve the conserved diversity, but from a certain amount of markers the advantage of increasing the number of markers is expected to be small. In a study of Gómez-Romano et al. (2012), the expected benefit from increasing the number of SNP markers over 1000 per Morgan (comparable to 30,000 SNP markers for cattle) is small. In our study the difference in overall genetic diversity based on pedigree data or SNP markers (42,000) was small (Chapter 3), indicating that high density SNP data or whole genome sequence data may not be of additional value for estimation and conservation of overall neutral genetic diversity within breeds. And also when we want to conserve a limited number of alleles, high density SNP data is not necessarily needed, because it will be enough to genotype animals for a small number of SNP markers. However, when we subsequently want to conserve the genetic diversity around these loci and at the rest of the genome, high density SNP data or whole genome sequence data will be useful. With an increasing number of SNPs, we have more possibilities to conserve specific alleles and at the same time the genetic diversity at the region around these alleles. In Chapter 5 in this thesis, we found that the loss of genetic diversity around a prioritized single allele can be substantial, emphasizing the importance of identifying and conserving all alleles around a targeted allele.

With whole genome sequence data, we have even more possibilities to identify and conserve alleles across the genome. Sequence data is expected to contain all causal polymorphisms (Meuwissen, 2010), and we can identify all alleles over the genome including rare variants with low minor allele frequency (Li et al., 2011). With SNP data this is not possible, because SNP markers with low minor allele frequency are less likely to be discovered or not selected because of the chance of genotype errors, while those ignored SNP markers might be important to identify rare alleles (Helyar

et al., 2011). With sequence data we cannot miss important rare alleles, because we are able to observe all variation over the genome. In addition, with sequence data we are able to estimate genetic diversity for each breed, including smaller (local) breeds. Genetic diversity estimation with SNP markers might not always be accurate for those breeds, as SNP arrays contain SNP markers that are selected because they are segregating in a few breeds (often commercial breeds) that were used to develop the SNP array. Therefore, using those SNP arrays may lead to bias in estimated genetic diversity across other breeds (Albrechtsen et al., 2010; Groenen et al., 2011). We can also use sequence data to identify other variants than SNPs, for example copy number variation (Liu et al., 2010; Hou et al., 2011), which can play an important role in identifying differences between breeds and prioritize genes that influence traits or diseases. Genome wide association studies (GWAS) are already a standard method for human disease gene discovery (Cantor et al., 2010). This is an emerging area of research in human genetics, and within a few years we expect this will become a hot topic in animal genetics as well. Part of the coming studies will be focused on detection of interesting alleles, which might be important to conserve for the future.

When we are able to sequence a population and identify all rare alleles, we can theoretically conserve all current genetic diversity, however, it is not realistic to conserve all rare alleles in a population. The number of rare alleles in a population is large, and it will be impossible to conserve all rare alleles within a limited number of animals. Therefore, it is the question if sequence data is of additional value compared to SNP data. Another discussion point is that it is difficult to conserve many rare alleles when we do not know which genes and functions are underlying the alleles. One can say that, as long as the functions of underlying alleles are unknown, we have to conserve as much genetic diversity as possible, including rare alleles. But then we might also conserve rare alleles that represent unfavorable traits, while we would rather eliminate these alleles instead of keeping them in the population. In the future, when we have more information about the position of favorable and unfavorable genes at the genome, we might use sequence data to identify these alleles in a population and subsequently use this information for prioritization of animals. In the next chapter about practical implications of SNP markers for conservation, we further discuss the possibilities of using sequence data for prioritization of animals.

On the one hand, sequence data can give us a lot of information about the genetic diversity in detail across the genome within populations. Sequence data might be necessary in the future when we want to conserve or eliminate specific genes, which might be missed when we use SNP data. On the other hand, it is the question if these

advantages outweigh the costs and efforts for sequencing, data analysis and labor. Sequencing results in a large amount of information, much more than we can possibly analyze. When we want to use this information for prioritization of animals, we have to decide what we want to conserve. In addition, at this moment the costs for sequencing will probably be too high for gene banks to make it affordable, although we expect costs will become much lower in the future and sequencing may become interesting for gene banks. Before we can use sequence data for genetic diversity conservation, more research is necessary to identify genes across the genome and find out their biological effects on traits. Additionally, practical implications for genetic diversity conservation with sequence data have to be developed.

6.4 Practical implications of the use of SNP markers for conservation

In this thesis we have shown that SNP markers can be used to identify genetic diversity in detail across the genome, and that we can use this information to reduce the loss of genetic diversity. A next step is to translate this knowledge into practical implications of SNP markers for conservation purposes. To support *in situ* conservation of populations, to make better decisions in *ex situ* conservation in a gene bank and subsequently to screen and utilize stored genetic material to help live populations when they have genetic problems or when they are in danger of extinction. However, in practice it can be difficult to collect, analyze and use large amounts of SNP and sequence data. In this chapter, I will describe some practical situations to illustrate that genomic information can be more or less useful in different situations, and that the decision to assemble and use different kinds of information must be guided by the question at hand.

Efficient conservation of genetic diversity

The practical use of SNP data for conservation of genetic diversity will be different for each situation, and not in every situation SNP data is needed or of advantage. When only a small number of animals of a breed is available for conservation in a gene bank, we do not have to make a choice which animals to conserve and we can simply store all animals in the gene bank. Then we do not need any genetic information. In situations in which we do need to prioritize animals, we have to observe the genetic diversity to make the best selection. This can be relatively simple, for example when we want to safeguard the several lines within a population in the gene bank. To be sure that from each line animals are stored in the gene bank, we need to discriminate

between the different lines in the population. For this purpose, pedigree information or microsatellites will be sufficient, as both can be used to observe the structure within a population (Vicente et al., 2008; Dalvit et al., 2009; Bouquet et al., 2011). A more difficult situation is when we need to discriminate between animals. When there is the possibility to store more animals per line or breed in a gene bank, we will need to observe the relatedness between animals to find out which animals have to be prioritized to maximize the conserved genetic diversity. Prioritization of animals can be done with pedigree data in combination with optimal contribution selection, which was proved to be a good method (Chapter 3). With SNP data we can improve prioritization (Chapter 3 and 4), however, in practice we have to consider whether the advantage of SNP data is large enough to outweigh the extra costs and effort. For example, when due to the extra costs for genotyping another important breed cannot be stored in the gene bank, it might be a better idea to use only pedigree data to prioritize animals and store all breeds that need to be conserved.

Conservation of specific genetic diversity

For both *in situ* and *ex situ* conservation, SNP markers can be used to identify and conserve specific alleles or haplotypes. Part of the animals in a population may harbor these specific alleles or haplotypes that are completely missing in the rest of the population or in other breeds. They may represent important traits that we want to safeguard for the future. Especially alleles with adaptive effects may be worth keeping for the future (Toro and Maki-Tanila, 2007), or an allele that is representing a favorable trait like polledness (Medugorac et al., 2012). For *in situ* conservation, we can use SNP markers to select animals with important alleles to produce the next generation, in order to maintain these alleles in the population. In particular for small populations this can be of importance, because they often have higher rates of inbreeding and therefore a higher chance of alleles being lost due to drift (Lacy, 1987). For *ex situ* conservation, we can use SNP markers to prioritize animals with important alleles to store in a gene bank. In this way we can safeguard these important alleles for the future, in case this allele will get lost in the live population (Oldenbroek, 2007).

However, we have to keep in mind that with SNP markers we cannot identify all alleles over the genome, in particular the rare alleles with a very low allele frequency. Therefore, sequence data can be useful for conservation of specific alleles, as with sequence data we can uncover all the variation that is present for each specific allele in a population. This is especially of importance for small breeds with high inbreeding rates, because they have a higher chance of losing rare alleles. When a number of animals in a population harbors an important allele, we can use sequence data to

identify the animals with this allele, and subsequently conserve genetic material of these animals in the gene bank. However, as whole genome sequencing results in a lot of data and substantial costs, it might be practical to combine pedigree and marker information and prioritize animals step by step. When pedigree information is already available, we can make a first selection of the least related animals by using optimal contribution selection based on pedigree information. Then we can genotype these animals with a high density SNP chip and subsequently use optimal contribution selection based on genomic relationships to select the animals that contribute the most to the genetic diversity. Those selected animals can be sequenced, and we can identify the animals that contain the favorable allele to be stored in the gene bank. In case we already know the position of the allele that we want to conserve, we can also sequence only that part of the genome we are interested in. Because we can lose substantial amounts of genetic diversity when we conserve a single allele (see Chapter 5), we can use SNP markers to monitor and prevent loss of diversity when selecting animals for the target allele with sequence data. When a larger budget is available or when costs will be very low in the future, we might sequence all animals that were selected based on pedigree information, and combine the selection for the target allele with reducing the loss of genetic diversity around the target allele. This can be done with optimal contribution selection, in which we use two constraints: one to conserve the target allele according to a certain allele frequency, and one to maximize the genetic diversity at the regions around the target allele.

Next to conserving one specific allele or haplotype, conservation of genetic diversity more generally also aims at conserving functional variation of a certain trait that is distributed over several parts of the genome. One breed can harbor genetic variation that is not found in other breeds, and it might be important to conserve this variation within the breed. For example, it might be a good idea to safeguard genetic material of cattle breeds that have a milk fatty acid composition that is favorable for human health (Maurice-Van Eijndhoven et al., 2011). These breeds might be unique for such a trait in comparison with other breeds, and worth saving for the future. Such production traits are often affected by large numbers of QTL (Stoop et al., 2009), and therefore we need to conserve genetic variation across the whole genome. High density SNP data can be very useful to characterize different breeds, and to map the breed specific genetic variation (The Bovine HapMap Consortium, 2009). Additionally, high density SNP data (i.e. >300,000 SNP in cattle) is predicted to be useful to find loci or haplotypes that have similar effects across breeds (De Roos et al., 2008; De Roos et al., 2009). This potentially enables to perform genomic selection in small breeds, aided by using multi-breed reference populations that include the mainstream

breeds as well. In the next few years, a lot of research will be geared towards this area. While this research effort may be focused on the possibilities for across-breed genomic predictions, it will at the same time also generate insight in the (genomic) differences between traits in different breeds. That information is potentially useful for future conservation decisions.

Conservation of breeds

Another implication of SNP markers for conservation of genetic diversity is the possibility to conserve genetic diversity for a group of breeds. Normally gene banks conserve genetic diversity for each breed separately, and often for each breed the same number of animals is prioritized for conservation. However, some breeds might be more important to conserve than others. For example, it is more interesting to conserve a breed with unique genetic diversity than a breed that is very similar to other breeds, and therefore we might need to store more genetic material of the unique breeds. Eding et al. (2002) mentioned this genetic overlap between breeds, and estimated the contribution of each breed to a core set. Conservation according to these contributions means that more genetic material is conserved of the unique breeds that received a higher contribution, and less from breeds that are more similar to other breeds. This results in a more efficient conservation of genetic diversity, which can be very important for gene banks because of their often limited budget. To put this method into practice for conservation of breeds in a gene bank, we can describe the following practical example. We assume that pedigree information is available. First, a core set can be formed of animals from commercial breeds, in which the overlap of genetic diversity is minimized. This can be done by using pedigree relatedness. The next step is to make a first selection of animals per breed with optimal contribution selection based on pedigree relationships. Subsequently, the selected animals can be genotyped, and we can use the method of Eding (2002) to estimate the contribution of each breed to the core set. The estimated contributions represent the uniqueness of each breed and of each animal within a breed, and by prioritizing animals based on these contributions we can conserve this uniqueness across breeds.

Efficient utilization of gene bank material

Genetic material that is stored in a gene bank can be used to help a population by increasing its genetic improvement, or saving it from genetic problems or extinction (Gandini and Oldenbroek, 2007). Before using stored genetic material from a gene bank, it is important to have a good knowledge on genetic diversity stored in the gene bank. SNP genotyping and even genome sequencing will be cheap in the future.

It becomes feasible to use SNP markers to screen current gene bank collections for genetic diversity. In that way it is possible to find out if collections harbor unique genetic material, and whether the collections lack genetic diversity at specific parts of the genome. Most genetic material in gene banks is collected based on pedigree information or on specific phenotypes. In some cases blood groups or microsatellites have been used to increase genetic diversity in the gene bank. We now have much better tools to characterize genetic material stored in gene banks. Instead of using SNP data to screen gene bank collections, it might soon also be cost effective to use whole genome sequencing. Sequence information will allow screening of the gene bank for rare variants.

Genetic characterization should be applied to ensure that gene banks are a good representation of the genetic diversity. Furthermore, genetic characterization can be used to select animals from the current population that should be added to the gene bank in order to maximize the stored genetic diversity. Animals can be identified in the current population that harbor genetic diversity which is not well represented in the existing gene bank. Genetic material of these animals can be added to the gene bank. In that way, gene bank collections can be improved. By genotyping the genetic material in the gene bank with a high density SNP chip, the genetic diversity can be estimated for not only the entire genome but also for specific regions of the genome. For adequate selection of specific regions, knowledge on the function of these regions is required. Recent studies on genomic diversity have identified areas in the genome that have been under selection, so-called selective sweeps (The Bovine HapMap Consortium, 2009; Rubin et al., 2010). This knowledge can be used to select regions that should be prioritized in gene banks.

It is impossible to conserve all genetic diversity. One scenario is that a gene bank wants to conserve as many breeds a possible, and because of a limited budget only a small number of animals per breed can be stored in the gene bank. In that situation it is of importance to prioritize animals in such way that as much genetic diversity as possible is conserved. In case of a deleterious mutation, conservation of genetic diversity is not desired. It also demonstrates that the value of additional SNP or even sequence information depends on our knowledge on the function of the genome.

New molecular methods provide a wealth of information that can be used in the conservation of genetic diversity in livestock. In this thesis possibilities of SNP markers are explored. Indeed new opportunities open up, but in practice these possibilities should always be balanced against costs and efforts needed to apply these methods. Research in this thesis revealed that genomic information can be effectively used not only to compare breeds for conservation of genetic diversity

over breeds, but also for screening to support conservation of genetic diversity within breeds.

References

A

Ajmone-Marsan, P., R. Negrini, P. Crepaldi, E. Milanesi, C. Gorni, A. Valentini, and M. Cicogna (2001). Assessing genetic diversity in Italian goat populations using AFLP (R) markers. Animal Genetics 32, 281-288.

Albrechtsen, A., F.C. Nielsen, and R. Nielsen (2010). Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Molecular Biology And Evolution 27, 2534-2547.

Avendaño, S., B. Villanueva, and J.A. Woolliams (2003). Expected increases in genetic merit from using optimized contributions in two livestock populations of beef cattle and sheep. Journal of Animal Science 81, 2964-2975.

B

Banos, G. and M.P. Coffey (2010). Short communication: Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows. Journal of Dairy Science 93, 2775-2778.

Bastiaansen, J.W.M., A. Coster, M.P.L. Calus, J.A.M.v. Arendonk, and H. Bovenhuis (2012). Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genetics Selection Evolution 44.

Baumung, R. and J. Solkner (2003). Pedigree and marker information requirements to monitor genetic variability. Genetics Selection Evolution 35, 369-383.

Beerda, B., W. Ouweltjes, L.B.J. Sebek, J.J. Windig, and R.F. Veerkamp (2007). Effects of genotype by environment interactions on milk yield, energy balance, and protein balance. Journal of Dairy Science 90, 219-228.

Bijma, P. (2012). Long-term genomic improvement – new challenges for population genetics. Journal Of Animal Breeding And Genetics 129, 1-2.

Bömcke, E. (2011). New method to combine molecular and pedigree relationships. Journal of Animal Science 89, 972-978.

Bouquet, A., E. Venot, D. Laloe, F. Forabosco, A. Fogh, T. Pabiou, K. Moore, J.A. Eriksson, G. Renand, and F. Phocas (2011). Genetic structure of the European Charolais and Limousin cattle metapopulations using pedigree analyses. Journal of Animal Science 89, 1719-1730.

Buch, L.H., M.K. Sørensen, P. Berg, L.D. Pedersen, and A.C. Sørensen (2012). Genomic selection strategies in dairy cattle: Strong positive interaction between use of genotypic information and intensive use of young bulls on genetic gain. Journal Of Animal Breeding And Genetics 129, 138-151.

Buys, C. (1990). Blood group frequencies in rare breeds of cattle. In: L. Alderson (ed.), Genetic conservation of domestic livestock. CAB International, Oxon, UK, 203-205.

C

Caballero, A. (1995). On the effective size of populations with seperate sexes, with particular reference to sex-linked genes. Genetics 139, 1007-1011.

Calus, M.P.L., T.H.E. Meuwissen, A.P.W. de Roos, and R.F. Veerkamp (2008). Accuracy of genomic selection using different methods to define haplotypes. Genetics 178, 553-561.

Canon, J., P. Alexandrino, I. Bessa, C. Carleos, Y. Carretero, S. Dunner, N. Ferran, D. Garcia, J. Jordana, D. Laloe, A. Pereira, A. Sanchez, and K. Moazami-Goudarzi (2001). Genetic diversity measures of local European beef cattle breeds for conservation purposes. Genetics Selection Evolution 33, 311-332.

Cantor, R.M., K. Lange, and J.S. Sinsheimer (2010). Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. The American Journal of Human Genetics 86, 6-22.

Carothers, A.D., I. Rudan, I. Kolcic, O. Polasek, C. Hayward, A.F. Wright, H. Campbell, P. Teague, N.D. Hastie, and J.L. Weber (2006). Estimating human inbreeding coefficients: Comparison of genealogical and marker heterozygosity approaches. Annals of Human Genetics 70, 666-676.

D

Daetwyler, H.D., F.S. Schenkel, and J.A.B. Robinson (2006). Relationship of multilocus homozygosity and inbreeding in Canadian Holstein sires. Canadian Journal of Animal Science 86, 578-579.

Daetwyler, H.D., B. Villanueva, P. Bijma, and J.A. Woolliams (2007). Inbreeding in genome-wide selection. Journal Of Animal Breeding And Genetics 124, 369-376.

Dalvit, C., M. De Marchi, R. Dal Zotto, E. Zanetti, T. Meuwissen, and M. Cassandro (2008). Genetic characterization of the Burlina cattle breed using microsatellites markers. Journal Of Animal Breeding And Genetics 125, 137-144.

Dalvit, C., M. De Marchi, E. Zanetti, and M. Cassandro (2009). Genetic variation and population structure of Italian native sheep breeds undergoing in situ conservation. Journal of Animal Science 87, 3837-3844.

Danchin-Burge, C., S.J. Hiemstra, and H. Blackburn (2011). Ex situ conservation of Holstein-Friesian cattle: Comparing the Dutch, French, and US germplasm collections. Journal of Dairy Science 94, 4100-8.

De Roos, A.P.W., B.J. Hayes, R.J. Spelman, and M.E. Goddard (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179, 1503-1512.

De Roos, A.P.W., B.J. Hayes, and M.E. Goddard (2009). Reliability of Genomic Predictions Across Multiple Populations. Genetics 183, 1545-1553.

De Roos, A.P.W., C. Schrooten, R.F. Veerkamp, and J.A.M. van Arendonk (2011). Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. Journal of Dairy Science 94, 1559-1567.

Dekkers, J.C.M. and F. Hospital (2002). The use of molecular genetics in the improvement of agricultural populations. Nature Reviews Genetics 3, 22-32.

Duclos, D. and S.J. Hiemstra (2010). State of local cattle breeds in Europe. In: S. J. Hiemstra, Y. De Haas, A. Maki-Tanila, and G. Gandini (eds.), Local cattle breeds in Europe. Wageningen Academic Publishers, Wageningen, 40-57.

E

Eding, H. and T.H.E. Meuwissen (2001). Marker-based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118, 141-159.

Eding, H., R. Crooijmans, M.A.M. Groenen, and T.H.E. Meuwissen (2002). Assessing the contribution of breeds to genetic diversity in conservation schemes. Genetics Selection Evolution 34, 613-633.

Efron, B. and R.B. Tibshirani (1993). An Introduction to the Bootstrap, Chapman and Hall, New York.

Ellegren, H. (2009). The different levels of genetic diversity in sex chromosomes and autosomes. Trends in Genetics 25, 278-284.

Engelsma, K.A., M.P.L. Calus, P. Bijma, and J.J. Windig (2010). Estimating genetic diversity across the neutral genome with the use of dense marker maps. Genetics Selection Evolution 42.

Engelsma, K.A., R.F. Veerkamp, M.P.L. Calus, and J.J. Windig (2011). Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information. Journal Of Animal Breeding And Genetics 128, 473-481.

Engelsma, K.A., R.F. Veerkamp, M.P.L. Calus, P. Bijma, and J.J. Windig (2012). Pedigree- and marker-based methods in the estimation of genetic diversity in small groups of Holstein cattle. Journal Of Animal Breeding And Genetics 129, 195-205.

F

Fabre Technology Platform (2011). Strategic Research Agenda 2011.

Fabuel, E., C. Barragan, L. Silio, M.C. Rodriguez, and M.A. Toro (2004). Analysis of genetic diversity and conservation priorities in Iberian pigs based on microsatellite markers. Heredity 93, 104-113.

Falconer, D.S. and T.F.C. Mackay (1996). Introduction to Quantitative Genetics, Longman Group, Essex, UK.

FAO (2007a). Global Plan of Action for Animal Genetic Resources and the Interlaken Declaration.

FAO (2007b). Livestock at risk. http://www.fao.org/ag/againfo/programmes/en/genetics/ITC_press_global.html.

FAO (2009). Status and trends report on animal genetic resources – 2008. ftp://ftp.fao.org/docrep/fao/meeting/016/ak220e.pdf.

Fernandez, J., B. Villanueva, R. Pong-Wong, and M.A. Toro (2005). Efficiency of the use of pedigree and molecular marker information in conservation programs. Genetics 170, 1313-1321.

Fernandez, J., T. Roughsedge, J.A. Woolliams, and B. Villanueva (2006). Optimization of the sampling strategy for establishing a gene bank: storing PrP alleles following a scrapie eradication plan as a case study. Animal Science 82, 813-821.

Fernandez, J., M.A. Toro, and A. Caballero (2008). Management of subdivided populations in conservation programs: Development of a novel dynamic system. Genetics 179, 683-692.

Fernandez, J., T.H.E. Meuwissen, M.A. Toro, and A. Maki-Tanila (2011). Management of genetic diversity in small farm animal populations. Animal 5, 1684-1698.

Fernando, R.L. and M. Grossman (1989). Marker assisted selection using best linear unbiased prediction. Genetics Selection Evolution 21, 467-477.

Flori, L., S. Fritz, F. Jaffrezic, M. Boussaha, I. Gut, S. Heath, J.L. Foulley, and M. Gautier (2009). The Genome Response to Artificial Selection: A Case Study in Dairy Cattle. Plos One 4, e6595.

Flury, C., M. Tietze, and H. Simianer (2006). Epistatic kinship a new measure of genetic diversity for short-term phylogenetic structures - theoretical investigations. Journal of Animal Breeding and Genetics 123, 159-171.

Flury, C., M. Tapio, T. Sonstegard, C. Drogemuller, T. Leeb, H. Simianer, O. Hanotte, and S. Rieder (2010). Effective population size of an indigenous Swiss cattle breed estimated from linkage disequilibrium. Journal Of Animal Breeding And Genetics 127, 339-347.

Freeman, A.R., D.G. Bradley, S. Nagda, J.P. Gibson, and O. Hanotte (2006). Combination of multiple microsatellite data sets to investigate genetic diversity and admixture of domestic cattle. Animal Genetics 37, 1-9.

G

Gandini, G.C. and J.K. Oldenbroek (1999). Choosing the conservation strategy. In: J. K. Oldenbroek (ed.), Genebanks and the management of farm animal genetic resources. Institute for Animal Science and Health, Lelystad, 11-33.

Gandini, G.C. and E. Villa (2003). Analysis of the cultural value of local livestock breeds: a methodology. Journal Of Animal Breeding And Genetics 120, 1-11.

Gandini, G.C. and J.K. Oldenbroek (2007). Strategies for moving from conservation to utilisation. In: K. Oldenbroek (ed.), Utilisation and conservation of farm animal genetic resources. Wageningen Academic Publishers, Wageningen, The Netherlands, 29-54.

Georges, M., M. Lathrop, Y. Bouquet, P. Hilbert, A. Marcotte, A. Schwers, J. Roupain, G. Vassart, and R. Hanset (1990). Linkage relationships among 20 genetic-markers in cattle - evidence for linkage between 2 pairs of blood-group systems - B-Z and S-F/V respectively. Animal Genetics 21, 95-105.

Goddard, M.E. (1992). Optimal effective population size for the global population of black and white dairy cattle. Journal of Dairy Science 75, 2902-2911.

Goddard, M.E. (2009). Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245-257.

Gómez-Romano, F., B. Villenueva, A. De Cara, and J. Fernández (2012). The benefit of using molecular coancestry in the management of populations under conservation and its dependency on effective population size and marker density. In: 4th International Conference on Quantitative Genetics. 2012.

Gourdine, J.L., A.C. Sorensen, and L. Rydhmer (2012). There is room for selection in a small local pig breed when using optimum contribution selection: A simulation study. Journal of Animal Science 90, 76-84.

Grapes, L., J.C.M. Dekkers, M.F. Rothschild, and R.L. Fernando (2004). Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci. Genetics 166, 1561-1570.

Groenen, M.A.M., H.J. Megens, Y. Zare, W.C. Warren, L.W. Hillier, R. Crooijmans, A. Vereijken, R. Okimoto, W.M. Muir, and H.H. Cheng (2011). The development and characterization of a 60K SNP chip for chicken. Bmc Genomics 12, 274.

Grundy, B., B. Villanueva, and J.A. Woolliams (1998). Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. Genetical Research 72, 159-168.

H

Habier, D., R.L. Fernando, and J.C.M. Dekkers (2007). The impact of genetic relationship information on genome-assisted breeding values. Genetics 177, 2389-2397.

Hagger, C. (2005). Estimates of genetic diversity in the brown cattle population of Switzerland obtained from pedigree information. Journal Of Animal Breeding And Genetics 122, 405-413.

Hayes, B.J. and M.E. Goddard (2008). Technical note: Prediction of breeding values using marker-derived relationship matrices. Journal of Animal Science 86, 2089-2092.

Hayes, B.J., S. Lien, H. Nilsen, H.G. Olsen, P. Berg, S. Maceachern, S. Potter, and T.H.E. Meuwissen (2008). The origin of selection signatures on bovine chromosome 6. Animal Genetics 39, 105-111.

Helyar, S.J., J. Hemmer-Hansen, D. Bekkevold, M.I. Taylor, R. Ogden, M.T. Limborg, A. Cariani, G.E. Maes, E. Diopere, G.R. Carvalho, and E.E. Nielsen (2011). Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Molecular Ecology Resources 11, 123-136.

Hiemstra, S.J. (2003). Guidelines for the constitution of national cryopreservation programmes for farm animals. In. Publication No. 1 European Region Focal Point For Animal Genetic Resources.

Hill, W.G. and A. Robertson (1968). Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38, 226-231.

Hill, W.G. (1993). Variation in genetic identity within kinships. Heredity 71, 652-653.

Hou, Y., G. Liu, D. Bickhart, M. Cardone, K. Wang, E.-s. Kim, L. Matukumalli, M. Ventura, J. Song, P. VanRaden, T. Sonstegard, and C. Van Tassell (2011). Genomic characteristics of cattle copy number variations. Bmc Genomics 12, 127.

J

Jacquard, A. (1983). Heritability - one word, 3 concepts. Biometrics 39, 465-477.

Jannink, J.L. (2010). Dynamics of long-term genomic selection. Genetics Selection Evolution 42, 35.

K

Kantanen, J., J. Vilkki, K. Elo, and A. Makitanila (1995). Random amplified polymorphic DNA in cattle and sheep - application for detecting genetic-variation. Animal Genetics 26, 315-320.

Kearney, J.F., E. Wall, B. Villanueva, and M.P. Coffey (2004). Inbreeding trends and application of optimized selection in the UK Holstein population. Journal of Dairy Science 87, 3503-3509.

Khatkar, M.S., F.W. Nicholas, A.R. Collins, K.R. Zenger, J. Al Cavanagh, W. Barris, R.D. Schnabel, J.F. Taylor, and H.W. Raadsma (2008). Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genomics 9, 187.

Kijas, J.W., D. Townley, B.P. Dalrymple, M.P. Heaton, J.F. Maddox, A. McGrath, P. Wilson, R.G. Ingersoll, R. McCulloch, S. McWilliam, D. Tang, J. McEwan, N. Cockett, V.H. Oddy, F.W. Nicholas, H. Raadsma, and C. International Sheep Genomics (2009). A Genome Wide Survey of SNP Variation Reveals the Genetic Structure of Sheep Breeds. Plos One 4, e4668.

Kim, E.S. and B.W. Kirkpatrick (2009). Linkage disequilibrium in the North American Holstein population. Animal Genetics 40, 279-288.

Kinghorn, B.P., R. Banks, C. Gondro, V.D. Kremer, S.A. Meszaros, S. Newman, R.K. Shepherd, R.D. Vagg, and J.H.J. van der Werf (2009). Strategies to Exploit Genetic Variation While Maintaining Diversity. In: J. H. J. van der Werf, H. U. Graser, R. Frankham, and C. Gondro (eds.), Adaptation and Fitness in Animal Populations - Evolutionary and Breeding Perspectives on Genetic Resource Management. Springer, The Netherlands, 191-200.

Koenig, S. and H. Simianer (2006). Approaches to the management of inbreeding and relationship in the German Holstein dairy cattle population. Livestock Science 103, 40-53.

L

Lacy, R.C. (1987). Loss of Genetic Diversity from Managed Populations: Interacting Effects of Drift, Mutation, Immigration, Selection, and Population Subdivision. Conservation Biology 1, 143-158.

Larsen, B. and K.M. Hansen (1986). Linkage analysis of loci controlling blood-groups and the rectovaginal constriction syndrome in Jersey cattle. Animal Genetics 17, 277-282.

Lenstra, J.A. (2006). Marker-assisted conservation of European cattle breeds: an evaluation. Animal Genetics 37, 475-481.

Li, L., Y. Li, S.R. Browning, B.L. Browning, A.J. Slater, X.Y. Kong, J.L. Aponte, V.E. Mooser, S.L. Chissoe, J.C. Whittaker, M.R. Nelson, and M.G. Ehm (2011). Performance of Genotype Imputation for Rare Variants Identified in Exons and Flanking Regions of Genes. Plos One 6, e24945.

Li, Y., H.N. Kadarmideen, and J.C.M. Dekkers (2008). Selection on multiple QTL with control of gene diversity and inbreeding for long-term benefit. Journal Of Animal Breeding And Genetics 125, 320-329.

Lin, B.Z., S. Sasazaki, and H. Mannen (2010). Genetic diversity and structure in Bos taurus and Bos indicus populations analyzed by SNP markers. Animal Science Journal 81, 281-289.

Liu, G.E., Y. Hou, B. Zhu, M.F. Cardone, L. Jiang, A. Cellamare, A. Mitra, L.J. Alexander, L.L. Coutinho, M.E. Dell'Aquila, L.C. Gasbarre, G. Lacalandra, R.W. Li, L.K. Matukumalli, D. Nonneman, L.C.d.A. Regitano, T.P.L. Smith, J. Song, T.S. Sonstegard, C.P. Van Tassell, M. Ventura, E.E. Eichler, T.G. McDaneld, and J.W. Keele (2010). Analysis of copy number variations among diverse cattle breeds. Genome Research 20, 693-703.

Lynch, M. (1988). Estimation of Relatedness by DNA Fingerprinting. Molecular Biology and Evolution 5, 584-599.

Lynch, M. and B.G. Milligan (1994). Analysis of population genetic-structure with RAPD markers. Molecular Ecology 3, 91-99.

Lynch, M. and K. Ritland (1999). Estimation of pairwise relatedness with molecular markers. Genetics 152, 1753-1766.

Lynn, D.J., A.R. Freeman, C. Murray, and D.G. Bradley (2005). A genomics approach to the detection of positive selection in cattle: Adaptive evolution of the T-cell and natural killer cell-surface protein CD2. Genetics 170, 1189-1196.

M

MacEachern, S., B. Hayes, J. McEwan, and M. Goddard (2009). An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. BMC Genomics 10.

Maudet, C., G. Luikart, and P. Taberlet (2002). Genetic diversity and assignment tests among seven French cattle breeds based on microsatellite DNA analysis. Journal of Animal Science 80, 942-950.

Maurice-Van Eijndhoven, M.H.T., S.J. Hiemstra, and M.P.L. Calus (2011). Short communication: Milk fat composition of 4 cattle breeds in the Netherlands. Journal of Dairy Science 94, 1021-1025.

Medugorac, I., D. Seichter, A. Graf, I. Russ, H. Blum, K.H. Göpel, S. Rothammer, M. Förster, and S. Krebs (2012). Bovine Polledness – An Autosomal Dominant Trait with Allelic Heterogeneity. Plos One 7, e39477.

Melka, M.G. and F. Schenkel (2010). Analysis of genetic diversity in four Canadian swine breeds using pedigree data. Canadian Journal of Animal Science 90, 331-340.

Meuwissen, T.H.E. and Z. Luo (1992). Computing inbreeding coefficients in large populations. Genetics Selection Evolution 24, 305-313.

Meuwissen, T.H.E. (1997). Maximizing the response of selection with a predefined rate of inbreeding. Journal of Animal Science 75, 934-940.

Meuwissen, T.H.E. and M.E. Goddard (2001). Prediction of identity by descent probabilities from marker-haplotypes. Genetics Selection Evolution 33, 605-634.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819-1829.

Meuwissen, T.H.E. (2002). GENCONT: an operational tool for controlling inbreeding in selection and conservation schemes. In: Proceedings of 7th World Congress on Genetics Applied to Livestock Production. France, 19-23 August 2002.

Meuwissen, T.H.E. and A.K. Sonesson (2004). Genotype-assisted optimum contribution selection to maximize selection response over a specified time period. Genetical Research 84, 109-116.

Meuwissen, T.H.E. (2009). Towards consensus on how to measure neutral genetic diversity? Journal Of Animal Breeding And Genetics 126, 333-334.

Meuwissen, T.H.E. (2010). Use of whole genome sequence data for QTL mapping and genomic selection. In: Proceedings of the 9th World Congress Genetics Applied Livestock Production Congress. Leipzig, Germany, 1-6 August 2010.

Mrode, R., J.F. Kearney, S. Biffani, M. Coffey, and F. Canavesi (2009). Genetic relationships between the Holstein cow populations of three European dairy countries. Journal of Dairy Science 92, 5760-5764.

Mucha, S. and J.J. Windig (2009). Effects of incomplete pedigree on genetic management of the Dutch Landrace goat. Journal Of Animal Breeding And Genetics 126, 250-256.

Muir, W.M., G.K.S. Wong, Y. Zhang, J. Wang, M.A.M. Groenen, R. Crooijmans, H.J. Megens, H. Zhang, R. Okimoto, A. Vereijken, A. Jungerius, G.A.A. Albers, C.T. Lawley, M.E. Delany, S. MacEachern, and H.H. Cheng (2008). Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. Proceedings Of The National Academy Of Sciences Of The United States Of America 105, 17312-17317.

N

Nagahata, H. (2004). Bovine leukocyte adhesion deficiency (BLAD): A review. Journal of Veterinary Medical Science 66, 1475-1482.

Nielsen, H.M., A.K. Sonesson, and T.H.E. Meuwissen (2011). Optimum contribution selection using traditional best linear unbiased prediction and genomic breeding values in aquaculture breeding schemes. Journal of Animal Science 89, 630-638.

O

Oldenbroek, J.K. (2007). Utilisation and conservation of farm animal genetic resources, Wageningen Academic Publishers, Wageningen, The Netherlands.

Oliehoek, P.A., J. Windig, J.A.M. Arendonk, and P. Bijma (2006). Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. Genetics 173, 483-496.

Oliehoek, P.A. and P. Bijma (2009). Effects of pedigree errors on the efficiency of conservation decisions. Genetics Selection Evolution 41.

Orr, N., W. Back, J. Gu, P. Leegwater, P. Govindarajan, J. Conroy, B. Ducro, J.A.M. Van Arendonk, D.E. MacHugh, S. Ennis, E.W. Hill, and P.A.J. Brama (2010). Genome-wide SNP association–based localization of a dwarfism gene in Friesian dwarf horses. Animal Genetics 41, 2-7.

P

Pedersen, L.D., A.C. Sørensen, and P. Berg (2010). Marker-assisted selection reduces expected inbreeding but can result in large effects of hitchhiking. Journal Of Animal Breeding And Genetics 127, 189-198.

Powell, J.E., P.M. Visscher, and M.E. Goddard (2010). Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics 11, 800-805.

Prasad, A., R.D. Schnabel, S.D. McKay, B. Murdoch, P. Stothard, D. Kolbehdari, Z. Wang, J.F. Taylor, and S.S. Moore (2008). Linkage disequilibrium and signatures of selection on chromosomes 19 and 29 in beef and dairy cattle. Animal Genetics 39, 597-605.

Prayaga, K.C. (2007). Genetic options to replace dehorning in beef cattle - a review. Australian Journal of Agricultural Research 58, 1-8.

Pryce, J.E. and H.D. Daetwyler (2012). Designing dairy cattle breeding schemes under genomic selection: a review of international research. Animal Production Science 52, 107-114.

R

Reed, D.H. and R. Frankham (2001). How closely correlated are molecular and quantitative measures of genetic variation? A meta-analysis. Evolution 55, 1095-1103.

Roughsedge, T., B. Villanueva, and J.A. Woolliams (2006). Determining the relationship between restorative potential and size of a gene bank to alleviate the risks inherent in a scrapie eradication breeding programme. Livestock Science 100, 231-241.

Rubin, C.J., M.C. Zody, J. Eriksson, J.R.S. Meadows, E. Sherwood, M.T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallbook, F. Besnier, O. Carlborg, B. Bed'hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, and L. Andersson (2010). Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464, 587-591.

S

Schaffner, S.F. (2004). The X chromosome in population genetics. Nature Reviews Genetics 5, 43-51.

Scheet, P. and M. Stephens (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics 78, 629-644.

Schlotterer, C. (2004). The evolution of molecular markers - just a matter of fashion? Nature Reviews Genetics 5, 63-69.

Schulman, N.F., G. Sahana, T. Iso-Touru, S.D. McKay, R.D. Schnabel, M.S. Lund, J.F. Taylor, J. Virta, and J.H. Vilkki (2011). Mapping of fertility traits in Finnish Ayrshire by genome-wide association analysis. Animal Genetics 42, 263-269.

Selvaggi, M., C. Dario, V. Peretti, F. Ciotola, D. Carnicella, and M. Dario (2010). Inbreeding depression in Leccese sheep. Small Ruminant Research 89, 42-46.

Silió, L., A. Fernández, A. Mercadé, P. Martin-Palomino, M.A. López, J. Rodrigáñez, and C. Ovilo (2010). Measuring inbreeding in a closed pig strain from high-density SNPs genotypes. In: Proceedings of the 9th World Congress Genetics Applied Livestock Production Congress. Leipzig, Germany, 1-6 August 2010.

Skaarud, A., J.A. Woolliams, and H.M. Gjoen (2011). Strategies for controlling inbreeding in fish breeding programs; an applied approach using optimum contribution (OC) procedures. Aquaculture 311, 110-114.

Smith, J.M. and J. Haigh (2007). The hitch-hiking effect of a favourable gene. Genetics Research 89, 391-403.

Sonesson, A.K. and T.H.E. Meuwissen (2000). Mating schemes for optimum contribution selection with constrained rates of inbreeding. Genetics Selection Evolution 32, 231-248.

Sonesson, A.K. and T.H.E. Meuwissen (2001). Minimization of rate of inbreeding for small populations with overlapping generations. Genetical Research 77, 285-292.

Sonesson, A.K., L.L.G. Janss, and T.H.E. Meuwissen (2003). Selection against genetic defects in conservation schemes while controlling inbreeding. Genetics Selection Evolution 35, 353-368.

Sonesson, A.K., J.A. Wooliams, and T.H.E. meuwissen (2010). Maximising genetic gain whilst controlling rates of genomic inbreeding using genomic optimum contribution selection. In: Proceedings of the 9th World Congress on Genetics Applied to Livestock Production. Leipzig, Germany, 1-6 August 2010.

Sonstegard, T.S., L. Ma, J.B. Cole, G.R. Wiggans, C.P. Van Tassell, G. Liu, B. Mariani, B. Crooker, P.M. Vanraden, and M. Silva (2008). Genomic signatures of artificial selection in U.S. Holstein cows. In: Proceedings of the 31st Conference of the International Society of Animal Genetics. Amsterdam, The Netherlands, 20-24 July 2008. ISAG Proc. Poster 2098, http://www.isag.org.uk/ISAG/all/2008_ISAG_Amsterdam_P2000.pdf Accessed March 11, 2009.

Sørensen, A.C., M.K. Sørensen, and P. Berg (2005). Inbreeding in Danish dairy cattle breeds. Journal of Dairy Science 88, 1865-1872.

Sørensen, M.K., A.C. Sørensen, R. Baumung, S. Borchersen, and P. Berg (2008). Optimal genetic contribution selection in Danish Holstein depends on pedigree quality. Livestock Science 118, 212-222.

Stella, A., P. Ajmone-Marsan, B. Lazzari, and P.J. Boettcher (2010). Identification of selection signatures in cattle breeds selected for dairy production. Genetics 185, 1451-1461.

Stoop, W.M., A. Schennink, M.H.P.W. Visker, E. Mullaart, J.A.M. van Arendonk, and H. Bovenhuis (2009). Genome-wide scan for bovine milk-fat composition. I. Quantitative trait loci for short- and medium-chain fatty acids. Journal of Dairy Science 92, 4664-4675.

Szoke, S., I. Komlosi, E. Korom, M. Ispany, and S. Mihok (2004). A statistical analysis of population variability in Bronze Turkey considering gene conservation. Archiv Fur Tierzucht-Archives of Animal Breeding 47, 377-385.

T

Taggart, J., A. Ferguson, and F.M. Mason (1981). Genetic variation in Irish populations of brown trout (Salmo trutta L.): electrophoretic analysis of allozymes. Comparative Biochemistry and Physiology Part B: Comparative Biochemistry 69, 393-412.

Tapio, M., M. Ozerov, I. Tapio, M.A. Toro, N. Marzanov, M. Cinkulov, G. Goncharenko, T. Kiselyova, M. Murawski, and J. Kantanen (2010). Microsatellite-based genetic diversity and population structure of domestic sheep in northern Eurasia. Bmc Genetics 11.

The Bovine HapMap Consortium (2009). Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 324, 528-532.

Thomas, S. (1990). The Curly Horse identification project of the CS fund conservancy (a case study). In: L. Alderson (ed.), Genetic conservation of domestic livestock. CAB International, Oxon, UK, 154-159.

Thompson, J.R., R.W. Everett, and N.L. Hammerschmidt (2000). Effects of inbreeding on production and survival in Holsteins. Journal of Dairy Science 83, 1856-1864.

Toro, M.A. and A. Caballero (2005). Characterization and conservation of genetic diversity in subdivided populations. Philosophical Transactions of the Royal Society B-Biological Sciences 360, 1367-1378.

Toro, M.A. and A. Maki-Tanila (2007). Genomics reveals domestication history and facilitates breed development. In: J. K. Oldenbroek (ed.), Utilisation and conservation of farm animal genetic resources. Wageningen Academic Publishers, Wageningen, The Netherlands, 75-102.

Toro, M.A., J. Fernandez, and A. Caballero (2009). Molecular characterization of breeds and its use in conservation. Livestock Science 120, 174-195.

V

Vanraden, P.M. (2007). Genomic Measures of Relationship and Inbreeding. Interbull Bulletin 37, 33-36.

Veerkamp, R.F., J.K. Oldenbroek, H.J. Van Der Gaast, and J.H.J. Van Der Werf (2000). Genetic correlation between days until start of luteal activity and milk yield, energy balance, and live weights. Journal of Dairy Science 83, 577-583.

Vicente, A.A., M.I. Carolino, M.C.O. Sousa, C. Ginja, F.S. Silva, A.M. Martinez, J.L. Vega-Pla, N. Carolino, and L.T. Gama (2008). Genetic diversity in native and commercial breeds of pigs in Portugal assessed by microsatellites. Journal of Animal Science 86, 2496-2507.

Vignal, A., D. Milan, M. SanCristobal, and A. Eggen (2002). A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution 34, 275-305.

Visscher, P.M., S.E. Medland, M.A.R. Ferreira, K.I. Morley, G. Zhu, B.K. Cornes, G.W. Montgomery, and N.G. Martin (2006). Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. PLoS Genet 2, e41.

W

Weigel, K.A. (2001). Controlling inbreeding in modern breeding programs. Journal of Dairy Science 84, E177-E184.

Weir, B.S., L.R. Cardon, A.D. Anderson, D.M. Nielsen, and W.G. Hill (2005). Measures of human population structure show heterogeneity among genomic regions. Genome Research 15, 1468-1476.

Wiggans, G.R., T.S. Sonstegard, P.M. Vanraden, L.K. Matukumalli, R.D. Schnabel, J.F. Taylor, F.S. Schenkel, and C.P. Van Tassell (2009). Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. Journal of Dairy Science 92, 3431-3436.

Windig, J.J. and T.H.E. Meuwissen (2004). Rapid haplotype reconstruction in pedigrees with dense marker maps. Journal of Animal Breeding and Genetics 121, 26-39.

Windig, J.J., H. Meuleman, and L. Kaal (2007). Selection for scrapie resistance and simultaneous restriction of inbreeding in the rare sheep breed "Mergellander". Preventive Veterinary Medicine 78, 161-171.

Windig, J.J., B. Beerda, and R.F. Veerkamp (2008). Relationship between milk progesterone profiles and genetic merit for milk production, milking frequency, and feeding regimen in dairy cattle. Journal of Dairy Science 91, 2874-2884.

Woolliams, J.A. and E.A. Mantysaari (1995). Genetic contributions of Finnish Ayrshire bulls over 4 generations. Animal Science 61, 177-187.

Woolliams, J.A. and M. Toro (2007). What is genetic diversity? In: J. K. Oldenbroek (ed.), Utilisation and conservation of farm animal genetic resources. Wageningen Academic Publishers, Wageningen, The Netherlands, 55-74.

Worley, K., J. Carey, A. Veitch, and D.W. Coltman (2006). Detecting the signature of selection on immune genes in highly structured populations of wild sheep (Ovis dalli). Molecular Ecology 15, 623-637.

Wray, N.R. and M.E. Goddard (1994). Increasing long-term response to selection. Genetics Selection Evolution 26, 431-451.

Z

Zenger, K.R., M.S. Khatkar, J.A.L. Cavanagh, R.J. Hawken, and H.W. Raadsma (2007). Genome-wide genetic diversity of Holstein Friesian cattle reveals new insights into Australian and global population variability, including impact of selection. Animal Genetics 38, 7-14.

English summary

English summary

The availability of dense SNP marker data has resulted in new opportunities to estimate genetic diversity within livestock breeds in more detail, and to improve prioritization of animals for conservation of genetic diversity. It is hypothesized that SNP markers can give a better estimation of the genetic diversity within breeds than pedigree information, for both the overall genetic diversity and the genetic diversity at specific genome regions. We also hypothesize that SNP markers can help improve the prioritization of animals in order to conserve genetic diversity within breeds, and especially to conserve genetic diversity at specific genome regions. However, little is known about how the genetic diversity varies over the genome, and what the differences are between pedigree and SNP based diversity estimates. Additionally, we do not know how much more genetic diversity can be conserved when we use SNP markers instead of pedigree information, and what the effects are of conservation that targets a specific region or locus only. The aim of this thesis was to explore the opportunities of SNP markers for estimation and conservation of genetic diversity within livestock breeds.

In **Chapter 2**, two different methods to estimate genetic diversity with SNP markers were compared in a simulation study. A population was set up with individuals containing a 1-Morgan chromosome with 1665 SNP markers, and from this one an additional population was produced with lower marker density i.e. 166 SNP markers. Genetic diversity was estimated either by IBD probabilities or heterozygosity, and compared to each other and the true genetic diversity. Genetic diversities estimated by IBD probabilities and by heterozygosity were positively correlated, and correlations with the true genetic diversity were quite similar for the simulated population with a high marker density, both for specific regions (r=0.19-0.20) and large regions (r=0.61-0.64) over the genome. For the population with a lower marker density, the correlation with the true genetic diversity turned out to be higher for the IBD-based genetic diversity. For a population with a lower marker density, the IBD-based method gave a better prediction of the genetic diversity, since variation and recombination between markers are missed with heterozygosity. When high density markers were used this advantage disappeared, and the two methods gave similar results.

Genetic diversity is often evaluated using pedigree information, but the availability of large numbers of SNP markers makes it possible to evaluate genetic diversity in more detail over the genome. In **Chapter 3**, genetic diversity estimation with SNP markers and pedigree information were compared for two small related groups of Holstein animals genotyped with the 50k SNP chip. Genetic diversity was estimated with

coefficient of kinship (pedigree) and expected heterozygosity (SNP). SNP-based genetic diversity was estimated genome-wide, per chromosome and for parts of the genome with 5-Mb sliding windows, and for the latter significance of difference between groups was determined by bootstrapping. Both pedigree- and SNP-based diversity indicated more diversity in one of the groups; 26 of the 30 chromosomes showed significantly more diversity for the same group, as did 25.9% of the chromosome regions. The results showed that even in small populations that are genetically close, differences in diversity can be detected. Pedigree- and SNP-based diversity gave comparable differences, but SNP-based diversity shows on which chromosome regions these differences are based. Especially in a closely related population, animals can be very similar to each other based on pedigree relatedness, and with SNP markers we can find those animals that harbor unique genetic diversity.
When pedigree or SNP information is used to prioritize animals for conservation, they might result in a different selection of animals and a different amount of conserved genetic diversity. In **Chapter 4**, the use of pedigree information and SNP markers for prioritization of animals for conservation of overall genetic diversity in a gene bank was investigated. From two Holstein populations that were genotyped with the 50k SNP chip, animals were prioritized with optimal contribution selection based on pedigree en SNP information. Consequences for genetic diversity were compared for both the overall genetic diversity and the genetic diversity at the chromosomal level. Optimal contribution selection resulted in a higher genetic diversity to be conserved compared to prioritization without optimal contributions. The overall conserved genetic diversity was somewhat higher for prioritization with SNP data, but differences were small. Differences in conserved genetic diversity were larger at the chromosome level, where selection with SNP data resulted in higher genetic diversity for most chromosomes, but at some chromosomes selection with pedigree information resulted in higher genetic diversity. This means that the chance of losing diversity at specific parts of the genome is somewhat smaller when we use SNP markers. To optimize conservation strategies, genomic information can help improve the selection of animals for conservation in those situations where pedigree information is unreliable or absent, or when we want to conserve diversity at specific genome regions.
Conservation of genetic diversity is often focused on the overall genetic diversity, but it might be favorable to conserve parts of the genome or even specific alleles. When animals are prioritized for one specific allele, for example for inclusion in a gene bank, this may result in the loss of diversity in other parts of the genome. In **Chapter 5**, SNP markers were used for prioritization of animals for a single allele in a gene bank, and the risk of losing genetic diversity was quantified. From a small

Holstein population, genotyped with the 50k SNP chip, animals were prioritized for a single allele by using optimal contribution selection. In order to do so, the optimal contribution method was extended with an extra constraint on the allele frequency of the target SNP marker. Results showed that elimination or fixation of alleles can result in substantial losses in genetic diversity around the targeted locus and also at the rest of the genome, depending on the allele frequency and the target frequency. Losses of genetic diversity around the target allele are the largest when the target frequency is very different from the current allele frequency. But it is also possible to conserve more genetic diversity by increasing low allele frequencies, like for example changing the allele frequency from 0.10/0.90 to 0.50/0.50.

From this PhD thesis, we can conclude that dense SNP data is a powerful tool for estimation and conservation of genetic diversity in livestock breeds. Although pedigree information gives a good representation of the overall genetic diversity, SNP markers can provide more detailed information about the genetic diversity over the genome. SNP markers can be used to identify differences in genetic diversity at the chromosomal level between animals, and subsequently conserve this genetic diversity using optimal contribution selection. Especially for small populations, SNP markers can play an important role in conservation of unique alleles, while simultaneously minimizing the loss of genetic diversity at the rest of the genome.

Nederlandse samenvatting

Nederlandse samenvatting

De komst van grote aantallen SNP merkers heeft geresulteerd in nieuwe mogelijkheden voor het gedetailleerd bepalen van genetische diversiteit in landbouwhuisdierrassen, en het verbeteren van prioritering van dieren voor het conserveren van genetische diversiteit. De verwachting is dat SNP merkers een betere voorspelling kunnen geven van de genetische diversiteit binnen rassen in vergelijking met stamboekinformatie, voor zowel de overall genetische diversiteit als de genetische diversiteit van specifieke chromosoomregio's. Tevens is de verwachting dat SNP merkers de prioritering van dieren voor de conservering van genetische diversiteit kan helpen verbeteren, en dan vooral voor het behoud van genetische diversiteit van specifieke regio's binnen het genoom. Er is echter weinig bekend over hoe genetische diversiteit varieert over het genoom, en wat het verschil is tussen genetische diversiteit gebaseerd op stamboekinformatie of SNP merkers. Daarnaast is het onbekend hoeveel meer genetische diversiteit we kunnen conserveren wanneer we SNP merkers gebruiken in plaats van stamboekinformatie, en wat het effect is van conservering van een specifieke regio of locus. Het doel van dit proefschrift was het verkennen van de mogelijkheden van SNP merkers voor bepaling en conservering van genetische diversiteit in landbouwhuisdierrassen.

In **hoofdstuk 2** zijn twee methoden ter bepaling van genetische diversiteit met behulp van SNP merkers vergeleken in een simulatiestudie. Een populatie werd gesimuleerd met voor ieder individu een 1-Morgan chromosoom met 1,665 SNP merkers. Vanuit deze populatie werd een tweede populatie opgezet met een lagere merkerdichtheid, met slechts 166 SNP merkers in totaal. Genetische diversiteit werd bepaald op basis van de kans op "IBD" (Identiek door afstamming) en op basis van heterozygotie. Deze werden vervolgens vergeleken met elkaar en met de werkelijke genetische diversiteit. Een positieve correlatie werd gevonden tussen de bepaling van genetische diversiteit met IBD en heterozygotie. In de populatie met hoge merkerdichtheid werden voor de twee toegepaste methoden vergelijkbare correlaties met de werkelijke genetische diversiteit gevonden, voor zowel de specifieke genoomregio's (0.19-0.20) als de grotere genoomregio's (0.61-0.64). In de populatie met lage merkerdichtheid werd voor de IBD methode een hogere correlatie met de werkelijke genetische diversiteit gevonden. Voor lagere merkerdichtheden is de IBD methode een betere voorspeller van de genetische diversiteit, omdat met heterozygotie variatie en recombinatie tussen merkers niet meegenomen wordt. Bij hogere merkerdichtheden verdwijnt dit voordeel en geven de twee methoden vergelijkbare resultaten.

Genetische diversiteit wordt veelal bepaald met behulp van stamboekinformatie, maar de komst van grote hoeveelheden SNP merkers maakt het mogelijk om genetische diversiteit in meer detail over het genoom te bepalen. In **hoofdstuk 3** is de bepaling van genetische diversiteit met SNP merkers en stamboekinformatie vergeleken voor twee kleine, sterk gerelateerde Holstein populaties, gegenotypeerd met de 50k SNP chip. Genetische diversiteit werd bepaald met de verwantschapscoëfficiënt op basis van stamboekinformatie en de verwachte heterozygotie op basis van SNP informatie. Genetische diversiteit op basis van SNP informatie werd bepaald over het gehele genoom, per chromosoom en voor delen van het genoom aan de hand van het voortschrijdend gemiddelde over 5-Mb grote stukken van het genoom. Voor de laatstgenoemde werd significantie van de verschillen bepaald met behulp van bootstrapping. Zowel stamboekinformatie als SNP informatie toonden een hogere genetische diversiteit in een van de twee groepen aan; 26 van de 30 chromosomen lieten een hogere genetische diversiteit zien voor dezelfde groep, en dit gold voor 25,9% van de chromosoomregio's. De resultaten laten zien dat, zelfs in kleine populaties waarin dieren sterk aan elkaar verwant zijn, verschillen in genetische diversiteit gedetecteerd kunnen worden. Stamboekinformatie en SNP merkers lieten vergelijkbare verschillen in genetische diversiteit zien, maar met SNP merkers kunnen we erachter komen op welke chromosoomregio's deze verschillen zich bevinden. Vooral in een sterk verwante populatie kunnen dieren erg op elkaar lijken gebaseerd op stamboekinformatie, terwijl met SNP merkers het mogelijk is om de dieren met unieke genetische diversiteit te vinden.

Het gebruik van stamboekinformatie of SNP merkers voor prioritering van dieren in een genenbank kan resulteren in een andere selectie van dieren en een verschil in de geconserveerde genetische diversiteit. In **hoofdstuk 4** is het gebruik van stamboekinformatie en SNP merkers voor prioritering van dieren voor conservering van genetische diversiteit in een genenbank verkend. Vanuit twee Holstein populaties, gegenotypeerd met de 50k SNP chip, werden dieren geprioriteerd met behulp van optimale contributie selectie gebaseerd op stamboekinformatie of op SNP merkers. De consequenties voor de genetische diversiteit in de selectie werden vergeleken voor zowel de overall genetische diversiteit als de genetische diversiteit op chromosoomniveau. Optimale contributie selectie resulteerde in een hogere geconserveerde genetische diversiteit dan selectie zonder optimale contributies. De overall geconserveerde genetische diversiteit was iets hoger wanneer SNP merkers werden gebruikt, maar de verschillen waren klein. De verschillen in geconserveerde genetische diversiteit waren groter op chromosoomniveau. Prioritering op basis van SNP merkers resulteerde in een

hogere genetische diversiteit voor de meeste chromosomen, hoewel voor sommige chromosomen prioritering op basis van stamboekinformatie resulteerde in een hogere genetische diversiteit. Dit betekent dat de kans op verlies van genetische diversiteit op specifieke delen van het genoom wat kleiner is wanneer we SNP merkers gebruiken voor prioritering van dieren. Om conserveringsstrategieën te optimaliseren kunnen we merkerinformatie gebruiken voor het verbeteren van prioritering van dieren voor conservering in het geval van onjuiste of onvolledige stamboekinformatie, of wanneer we genetische diversiteit willen behouden op specifieke delen van het genoom.

Conservering van genetische diversiteit is meestal gericht op de overall genetische diversiteit, maar soms is het gewenst om delen van het genoom of zelfs specifieke allelen te bewaren. Wanneer we bijvoorbeeld dieren prioriteren voor opname in een genenbank op basis van één specifiek allel, kan dit resulteren in verlies van genetische diversiteit op andere delen van het genoom. In **hoofdstuk 5** zijn SNP merkers gebruikt voor de prioritering van dieren voor een specifiek allel, en is gekeken naar de kans op verlies van genetische diversiteit. Vanuit een kleine Holstein populatie, gegenotypeerd met de 50k SNP chip, zijn dieren geprioriteerd voor opname in een genenbank op basis van een specifiek allel met behulp van optimale contributie selectie. Hiervoor is de optimale contributie methode aangepast door als extra beperking een allelfrequentie van de SNP merker die we willen bewaren toe te voegen. De resultaten laten zien dat eliminatie of fixatie van allelen kan resulteren in substantiële verliezen van genetische diversiteit rond het geconserveerde allel en ook op andere delen van het genoom, afhankelijk van de originele allelfrequentie in de populatie en de doelfrequentie in de genenbank. Verlies van genetische diversiteit rond het geconserveerde allel is het grootst wanneer de doelfrequentie erg verschillend is van de huidige allelfrequentie. Maar we kunnen ook juist meer genetische diversiteit bewaren door een lage allelfrequentie te verhogen, bijvoorbeeld een allelfrequentie van 0.10/0.90 verhogen naar 0.50/0.50.

Op basis van dit proefschrift kunnen we concluderen dat grote hoeveelheden SNP merkers een grote rol kunnen spelen in het bepalen en conserveren van genetische diversiteit in landbouwhuisdierrassen. Hoewel stamboekinformatie een goede weergave is van de overall genetische diversiteit, kunnen SNP merkers meer gedetailleerde informatie geven over hoe de genetische diversiteit verdeeld is over het genoom. Met SNP merkers kunnen we verschillen in genetische diversiteit tussen dieren detecteren op chromosoomniveau, en tevens deze genetische diversiteit zo optimaal mogelijk conserveren met behulp van optimale contributie selectie. Vooral in kleine populaties kunnen SNP merkers een belangrijke rol spelen

in het behoud van unieke allelen, en tegelijkertijd het verlies van genetische diversiteit op de rest van het genoom beperken.

Dankwoord

En dan is het zover. Het harde werken wordt beloond, mijn proefschrift is af! De afsluiting van een periode waarin ik veel mensen heb mogen ontmoeten en van hen heb mogen leren. Ik ben dankbaar voor wat zij in deze tijd voor mij betekend hebben!

Allereerst de begeleidingscommissie. Johan, als mijn promotor, dank voor jouw vertrouwen en de positieve kijk op het geheel in de tijden dat het wat moeizamer verliep. Ondanks dat we elkaar niet zo vaak zagen ben je altijd erg betrokken geweest bij het proces, en in de allerlaatste eindsprint gaf je mij het juiste duwtje. Ik waardeer het enorm! Piter, jij had wel door dat een teveel aan theorie mij niet zo paste. Je hebt mij hierin ondersteund, en mij gemotiveerd om door zure appels heen te bijten (ook al wilde ik het veel liever wat praktischer houden!). Jack, als dagelijks begeleider had je altijd tijd voor mij om vragen te beantwoorden of te discussiëren. De koers is nog wel eens gewijzigd, maar uiteindelijk hebben we een mooi stukje werk kunnen leveren! Sipke Joost, de discussies met jou over conservering van genetische diversiteit binnen de genenbank heb ik machtig interessant gevonden (helemaal omdat het veelal over praktische zaken ging!). Je was een steun voor mij wanneer ik weer eens in het dal zat, en daar ben ik je dankbaar voor! Mario, jij bent vaak mijn reddende engel geweest ☺. Met het rekenwerk heb je mij enorm geholpen, en geen probleem was te groot om opgelost te worden. En ook in de afronding was je altijd daar om te helpen. Bedankt!

Roel, als afdelingshoofd was procesbewaking jouw taak. Je hebt mij veel geleerd op gebied van focus, planning en doorzetten. Nou, doorgezet heb ik! Het was lang niet altijd makkelijk, maar het zijn waardevolle lessen voor mij geweest.

Al mijn afdelingsgenoten van de afdeling Genomica, dank jullie wel voor alle gezelligheid en de interesse in mijn onderzoek! Dirkjan and Arun, you started at the same time as PhD student in Lelystad, thank you for the nice 'aio-time' we had together! Fortunately for me, you completed your PhD project somewhat earlier, and you could help me out with all kinds of questions ☺. Rianne en Yvonne, 'jonge aanwas' van onze afdeling en mijn trouwe 'thee- en kamergenootjes' tijdens mijn laatste loodjes. Jullie hebben mij vaak opgefleurd met jullie heerlijke droge humor, en jullie waren een erg goede helpdesk/PA ☺. Over een paar jaar kom ik jullie aanmoedigen!

Mijn collega's binnen Wageningen UR Livestock Research, van de afdeling Milieu maar ook van Veehouderijsystemen, bedankt voor de fijne samenwerking het afgelopen jaar en de interesse in mijn onderzoek!

Bastiaan, mijn baas bij de afdeling Milieu en goede vriend, jij hebt mij altijd heel erg duidelijk gemaakt dat je in mij gelooft en mij waardeert om de kwaliteiten die ik heb. Dat heeft mij zoveel geholpen! Dank voor de kans om mijn proefschrift verder af te maken tijdens mijn werk bij Milieu. En natuurlijk voor jouw vertrouwen in mij, dat ik het kon ☺.

Myrthe en Marike, ik ben zo blij dat jullie naast mij zitten tijdens mijn verdediging! Wat had ik zonder jullie gemoeten in al die tijd. Jullie hebben zoveel 'gezeur en geklaag' van mij aangehoord, als er weer eens een i met een j verwisseld was rond de kerstdagen, of wanneer de computer eigenhandig besloot dat hij het niet meer aan kon. Met z'n drieën hebben we vele ups maar ook zeker downs meegemaakt, maar het heeft ons sterker dan ooit gemaakt. Onze liefde voor paarden maakt ons sterk verbonden, en zorgde voor een extra thema tijdens onze aio-tijd: rechtrichten. Dit wordt uiteraard gewoon gezamenlijk voortgezet! Marike, als mijn kamergenootje was het heerlijk om altijd mijn verhaal bij jou kwijt te kunnen, zowel werk gerelateerd als privé. Jij bent zo heerlijk nuchter en dat maakte dat ik de zaken beter kon relativeren. Dank je wel! Myrthe, jij hebt mij wel de meest wijze les geleerd: geniet van het leven, want het kan maar zo voorbij zijn. Wat ben ik blij dat je er bent, als vriendinnetje die altijd voor mij klaar staat. Dikke kus!

Han, ik heb jou leren kennen toen je voorzitter was van de Nederlandse Zoötechnische Vereniging en ik als (jong en vrouwelijk ☺) bestuurslid bij de vereniging kwam. Dat was een prachtige tijd waarin ik een mooi netwerk heb kunnen opbouwen. Op het moment dat ik in mijn diepste dip zat heb jij mij het juiste zetje gegeven!

Letty, bij jou in de winkel is toch een zekere basis gelegd, waar ik altijd profijt van zal hebben. En ik ben het nog steeds niet verleerd!

Lieve kennissen, vrienden, vriendinnen en paardrijvriendinnetjes, dank jullie wel voor jullie interesse in mijn onderzoek, de steuntjes in de rug en de nodige ontspanning die oh zo belangrijk is voor een aio!

Anneke en Paul, jullie hebben mij letterlijk door de laatste loodjes heen gesleurd. Anneke, jouw nuchtere instelling, doortastendheid en stappenplan (die blijkt multidisciplinair ☺) hebben ervoor gezorgd dat ik het overzicht hield en dat ik gefocust bleef op de juiste dingen. Jullie zijn er altijd voor mij, in goede en in slechte tijden, en dat waardeer ik enorm!

Arjan, je hebt gelijk, het komt inderdaad wel goed. Je hebt mij blij gemaakt in de tijd dat een hectische periode overging in afronding van veel dingen, dank je wel hiervoor!

Jack, mijn trouwe viervoeter, door jou was ik in staat om in de avonden alles even helemaal te vergeten, en mijn hoofd leeg te maken. Jij hebt de gave om mij een spiegel voor te houden, wat erg confronterend is maar zo ontzettend leerzaam. Ik hoop nog heel lang van je te kunnen genieten!

En tot slot, mama, papa en Lisa, jullie zijn van onschatbare waarde voor mij. Jullie staan altijd voor de volle 100% achter mij, in alles wat ik in mijn leven besluit te gaan doen. Lisa, wij zijn samen opgegroeid tussen de beesten en delen een voor ons belangrijke passie: paarden. Dank je wel dat je er altijd bent voor mij! Mama en papa, dat volste vertrouwen in mij en de warmte en liefde die jullie mij en Lisa hebben meegegeven in onze opvoeding hebben mij gemaakt tot wie ik nu ben. Papa, mijn liefde voor dieren en interesse in onderzoek komen niet van een vreemde. Jouw oneindige kennis heb ik altijd enorm bewonderd, en je hebt ons van jongs af aan zoveel geleerd. Ik ben ontzettend trots dat je samen met mij de voorkant van mijn boekje hebt ontworpen, met een prachtige foto! Mama, jouw sociale en organisatorische kwaliteiten heb je aan mij doorgegeven. Ondanks dat in het onderzoek veelal de 'blauwe' kwaliteiten in het voordeel zijn, ben ik maar wat blij met deze kwaliteiten. Wat jij voor mij doet, onvoorwaardelijk, is eigenlijk niet in woorden uit te drukken. Dank jullie wel!

Krista

Curriculum Vitae

Curriculum Vitae

Krista Anika Engelsma was born on December 15 1980 in Wageningen, The Netherlands. She was raised in Rhenen and in 1999 she graduated from high school Christelijk Lyceum Veenendaal and started with the BSc Animal Husbandry at HAS Den Bosch. In 2003 she graduated and started with the MSc Animal Sciences at Wageningen University, with the specializations 'Animal Breeding and Genetics' and 'Behavior and Welfare'. She performed one major thesis at the department Genomics at Wageningen UR Livestock Research, on the genetic diversity in 70 European cattle breeds using marker data. A second major thesis was performed at the Institute for Pig Genetics (IPG), on behavior in sows and the relation with mothering ability. In 2006 she started working at IPG as a research assistant, and in 2007 she started her PhD study at the Animal Breeding and Genetics group of Wageningen University, in collaboration with the department Genomics at Wageningen UR Livestock Research. The results of this research are described in this thesis. In August 2011 she started working as a researcher at the department Environment of Wageningen UR Livestock Research.

Curriculum Vitae

Krista Anika Engelsma werd geboren op 15 december 1980 te Wageningen, en groeide op in Rhenen. In 1999 behaalde ze haar HAVO diploma aan het Christelijk Lyceum Veenendaal, waarna ze begon aan de studie Dierhouderij aan de HAS Den Bosch. In 2003 haalde zij haar diploma, en begon aan de masters opleiding Dierwetenschappen aan Wageningen Universiteit, met de specialisaties 'Fokkerij en Genetica' en 'Gedrag en Welzijn'. Twee afstudeervakken werden uitgevoerd. De eerste bij Wageningen UR Livestock Research in Lelystad bij de afdeling Genomica, gericht op de bepaling van de genetische diversiteit in 70 Europese runderrassen met behulp van merker data. De tweede bij het Institute for Pig Genetics (IPG), gericht op gedrag van zeugen en de relatie met moedereigenschappen. In 2006 is ze afgestudeerd en begon ze als onderzoeksassistent bij IPG. In 2007 begon ze als promovenda bij de leerstoelgroep Fokkerij en Genetica van Wageningen Universiteit, in samenwerking met de afdeling Genomica van Wageningen UR Livestock Research. De resultaten van dit onderzoek zijn beschreven in dit proefschrift. In augustus 2011 is zij gaan werken als onderzoeker bij de afdeling Milieu van Wageningen UR Livestock Research.

List of publications

Peer reviewed articles

- Engelsma, K.A., R.F. Veerkamp, M.P.L. Calus, P. Bijma and J.J. Windig (2012) Pedigree- and marker-based methods in the estimation of genetic diversity in small groups of Holstein cattle. Journal of Animal Breeding and Genetics 129 (3): 195-205.
- Engelsma, K.A., R.F. Veerkamp, M.P.L. Calus and J.J. Windig (2011) Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information. Journal of Animal Breeding and Genetics 128 (6): 473-481.
- Engelsma, K.A., M.P.L. Calus, P. Bijma and J.J. Windig (2010) Estimating genetic diversity across the neutral genome with the use of dense marker maps. Genetics Selection Evolution 42: 12.
- Windig, J.J. and K.A. Engelsma (2010) Perspectives of genomics for genetic conservation of livestock. Conservation Genetics 11 (2): 635-641.

Papers submitted

- Engelsma, K.A., R.F. Veerkamp, M.P.L. Calus and J.J. Windig. Consequences for diversity when animals are prioritized for conservation using the whole genome or one specific allele. Submitted to Journal of Animal Breeding and Genetics.

Conference proceedings

- Engelsma, K.A. and J.J. Windig (2011) Genomics and selection of animals for a gene bank. In Book of Abstracts of the 62nd Annual Meeting of the European Association for Animal Production (EAAP), Stavanger, Norway, 29 August - 2 September 2011.
- Engelsma, K.A., Veerkamp, R.F., Calus, M.P.L. and Windig, J.J. (2010) Differences in genetic diversity in Holstein cattle with high and low genetic merit. In 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Germany, 1-6 August 2010.
- Engelsma, K.A., Veerkamp, R.F., Calus, M.P.L. and Windig, J.J. (2010) Comparison of genetic diversity between high and low genetic merit in a Holstein population. In 5th International Cattle Breeders Round Table 2010, Sønderborg, Denmark, 11-13 January 2010.
- Engelsma, K.A., Calus, M.P.L., Hiemstra S.J., Bijma, P. and Windig, J.J. (2009) Using dense marker maps to determine genetic diversity over the neutral genome. In 60th Annual meeting of the European Association for Animal Production (EAAP), Barcelona, Spain, 24-27 August 2009.

- Engelsma, K.A., Calus, M.P.L., Hiemstra, S.J., Bijma, P., van Arendonk, J.A.M. and Windig, J.J. (2008) Genome wide diversity: variation over a genome. In International symposium: New opportunities for conservation genetics with genome wide information, Wageningen, The Netherlands, 8 December 2008.
- Engelsma, K.A., Calus, M.P.L., Hiemstra, S.J., van Arendonk, J.A.M., Windig, J.J. (2008) A method to determine variation in genetic diversity across the genome using dense marker maps. In 59th Annual meeting of the European Association for Animal Production (EAAP), Vilnius, Lithuania, 23-27 August 2008.

Training and Supervision Plan

Training and Supervision Plan

Basic package (3 ECTS)

WIAS Introduction Course	2007
Course on philosophy of science and/or ethics	2009

Scientific exposure (16 ECTS)

International conferences

58th Annual Meeting of the EAAP, Dublin, Ireland, August 26-29	2007
59th Annual Meeting of the EAAP, Vilnius, Lithuania, August 24-27	2008
ESF Conservation Genetics conference, Trondheim, Norway, May 24-26	2009
60th Annual Meeting of the EAAP, Barcelona, Spain, August 24-27	2009
5th International Cattle Breeders Round Table, Sønderborg, Denmark, January 11-13	2010
9th WCGALP, Leipzig, Germany, August 1-7	2010

Seminars and workshops

WIAS seminar Participatory conservation of indigenous breeds of livestock, June 11	2008
International symposium New opportunities for conservation genetics with genome-wide information, December 8	2008
F&G connection days, Vught, The Netherlands, November 27-28	2008
WIAS Science Day, Wageningen, The Netherlands, March 12	2009
F&G connection days, Vught, The Netherlands, November 25-26	2010
WIAS Science Day, Wageningen, The Netherlands, January 28	2010
WIAS Science Day, Wageningen, The Netherlands, February 3	2011

Presentations

59th Annual Meeting of the EAAP (oral)	2008
International symposium New opportunities for conservation genetics with genome-wide information (oral)	2008
60th Annual Meeting of the EAAP (oral)	2009
5th International Cattle Breeders Round Table (oral)	2010
WIAS Science Day (poster)	2010
9th WCGALP (oral)	2010

In-depth studies (6 ECTS)

Disciplinary and interdisciplinary courses

QTL mapping, MAS, and genomic selection, Lelystad, The Netherlands, March 10-14	2008
QTL-MAS course, Wageningen, The Netherlands, April 20-21	2009
Quantitative genetics with a focus on "selection theory", Wageningen, The Netherlands, June 7-11	2010
Genomic selection in livestock, Wageningen, The Netherlands, June 27-July 1	2011

PhD students' discussion groups

Theme group Biodiversity	2007-2011

Professional skills support courses (6 ECTS)

PhD competence assessment	2007
Project and time management	2007
Techniques for writing and presenting a scientific paper	2008
Writing for academic publication	2010
Career perspectives	2010

Didactic skills training (1 ECTS)

Assisting Genetic Improvement of Livestock (GIL)	2008

Management skills training (14 ECTS)

Organization of seminars and courses

Organization international symposium New opportunities for conservation genetics with genome-wide information	2008

Membership of boards and committees

Member of Young ASG	2008
Secretary of WIAS Associated PhD Students (WAPS) council	2008
Chairman of WIAS Associated PhD Students (WAPS) council	2009
Member of WIAS Associated PhD Students (WAPS) council	2010

Education and training total: 46 ECTS

Colophon

Colophon

The printing of this thesis was supported by the Centre for Genetic Resources, The Netherlands, funded by the Ministry of Economic Affairs, Agriculture and Innovation, program “Kennisbasis Dier”, code KB-04-002-021.

The cover of this thesis was designed by Frans J. Engelsma and Krista A. Engelsma.

This thesis was printed by GVO drukkers & vormgevers B.V. | Ponsen & Looijen, Ede, The Netherlands.